

Faster and Efficient VM Migrations for Improving SLA and ROI in Cloud Infrastructures

Sujal Das, Michael Kagan, and Diego Crupnicoff (diego@mellanox.com)

Abstract— Cloud platforms and infrastructures host several independent users and applications on a shared resource pool with the ability to allocate computing power to users and applications on a per-demand basis. The use of server virtualization techniques for such platforms provides great flexibility with the ability to allocate users and applications per virtual machine (VM), ability to consolidate several virtual machines on the same physical server, to resize a virtual machine capacity as needed and being able to support migration of virtual machines across physical servers to meet SLA and capacity constraints. A key challenge for cloud providers is their ability to deliver highest service levels (in terms of availability and performance through SLAs) to their customers while minimizing operational costs. SLAs currently provided by cloud service providers are weighted toward lower cost commodity-like services and do not meet the expectations of enterprise mission critical workloads. Numerous research activities performed in this area indicate the critical importance of efficient VM migration capabilities in these cloud infrastructures to deliver higher service levels at lower operational costs. VM migration over high performance I/O can deliver significantly better VM migration efficiencies to enable higher service levels, improved SLA guarantees while boosting return on investment for cloud service providers.

Index Terms—Cloud, Migration, Service level, Virtual machine.

I. INTRODUCTION

VIRTUAL machines (VMs) are becoming increasingly valuable for resource consolidation and management, providing efficient and secure resource containers, along with desired application execution environments. Coupling those resource utilization benefits with VM migration capabilities enable dynamic, elastic data centers as required in the cloud computing paradigm. In cloud computing environments where the cloud infrastructure is shared by multiple customers and applications, the VMs and the capabilities they provide, therefore, are the key building blocks and lifeline of the cloud service provider's data center and customers who outsource computing to those data centers.

It is important to note that server virtualization technologies were not built ground up to meet these cloud infrastructure requirements. They evolved first with mainframes, then for desktops, then for servers with test and development as

primary applications. Subsequently, better VM management capabilities and VM migration features facilitated deployments within enterprise data centers. The evolution of cloud infrastructure, their critical dependence on VMs to deliver services and SLAs to customers is driving server virtualization feature and efficiency needs to new directions. The role of the network and its usage by VMs has evolved, from simple 10/100 Mbps connectivity to Gigabit Ethernet connectivity and now to higher bandwidth and converged I/O solutions like 10GigE and InfiniBand. The legacy of VM usage practices and the reliance on use of ultra low-cost commodity hardware tend to influence data center designers and researchers alike as they overlook the importance of I/O in cloud infrastructures that brings many new paradigms. This paper focuses on:

- SLA challenges in current cloud infrastructures
- Discusses different SLA maximization and expense minimization approaches by citing relevant research work conducted by others and how they lead to the need for efficient VM migration strategies
- Discusses VM migration expenses and the role of network I/O in VM migration
- Presents VM migration efficiencies over high performance, low latency networks by citing relevant research work conducted by others
- Concludes how higher service levels and guaranteed SLAs can be achieved while lowering expenses by using techniques discussed in the paper

II. SLA CHALLENGES IN CURRENT CLOUD INFRASTRUCTURES

Wikipedia [1] defines SLA as: “The SLA records a common understanding about services, priorities, responsibilities, guarantees and warranties. Each area of service scope should have the ‘level of service’ defined. The SLA may specify the levels of availability, serviceability, performance, operation, or other attributes of the service such as billing. The ‘level of service’ can also be specified as ‘target’ and ‘minimum’, which allows customers to informed what to expect (the minimum), whilst providing a measurable (average) target value that shows the level of organization performance. In some contracts penalties may be agreed in the case of non compliance of the SLA (but see ‘internal’ customers below).It is important to note that the ‘agreement’ relates to the services the customer receives, and not how the service provider delivers that service.”

For enterprises that rely on its own data centers, SLA is guaranteed by the internal IT resources that have to ensure there is no downtime that causes loss of business and revenues. When such enterprises consider outsourcing computing services to a third party, such as cloud service provider, the equation of accountability changes considerably. The equation changes in that what is considered a given when using an internal IT infrastructure becomes variable and entails new tasks when using an external and shared cloud IT infrastructure. For example, the following definitions, documentations, and accountability questions arise:

- Defining outages
- How does a customer prove an outage to get credit?
- How does the credit get applied?
- Is SLA tied to application level performance or just availability of server hardware resources?

Today, one of the main contention points in negotiating a cloud services SLA is around the outage credits and how they are applied. Does the customer get a reimbursement for the lost services or is the SLA applied to a future credit? The test of a great SLA is one that gives a customer a direct reimbursement for lost services. Another area of difficulty when negotiating an SLA is defining the SLA outages. And at the end, the most important question is: Does the credit for outages compensate adequately for lost business? The only satisfactory answer to this question is no outages at all through incorporation of performance and five 9s availability criteria in SLAs. Customers run varied kinds of applications and performance of such applications should be tied to SLAs. For example, minimum jobs per second achievable per application with size of VMs purchased (through amount of CPU, memory, network and storage resources) should be a SLA criteria with failure to meet such requirements categorized as an outage requiring credit. We compare the most dominant cloud service provider – Amazon Web Services (AWS) EC2 and their SLA practices [2] with a competitive offering from 3Tera [3].

A. Defining The Outage

AWS: In the AWS SLA EC2 agreement, Amazon claims a 99.95% SLA. A defined outage in AWS is very confusing at best. It means that a customer cannot launch a replacement instance within a 5 minute period while at least two availability zones within the same region are down. What this means is that if two out of three data centers are available and the customer still cannot launch and/or run any application on its EC2 server, it will not be defined as an outage. To further complicate the matter, AWS calculates their 99.95% availability based on the previous 365 days. If the customer doesn't have 365 prior days of service with AWS the prior days are calculated as 100% available. This means that if the customer is a new customer (say 2 months old), and a catastrophic event happens to hit two of the three US based data centers and the customer cannot start an instance for three days, then the customer would get a 10% credit for only one day's prorated costs for

EC2 services. The first two days would not be below the 12 month period 99.95% availability SLA. Also complicating the AWS EC2 SLA is the new reserve instances' upfront fees are not eligible for credits concerning outages.

3Tera: 3Tera claims to deliver the highest cloud SLA - 99.999 percent for their Virtual Private Datacenter (VPDC) customers. The customer does not have to define the outage. 3Tera will automatically detect and calculate outages. The AppLogic Cloud Computing Platform constantly monitors and reports the availability of the system and instantly alerts 3Tera's operations team of critical issues. The automatic recording of outages is considered the unprecedented feature of their SLA. While other cloud vendors require the customer to prove the outage times, 3Tera automates this process.

B. How Does A Customer Prove An Outage To Get Credit?

AWS: In order to receive a credit for a defined AWS EC2 outage, a customer has to capture, document, and send a request to Amazon to be processed. In other words, the onus is on the customer to prove the outage. AWS does not provide any interface or documentation to help the customer define their outages. Furthermore, Amazon requires the customer to document the region, all instance ids, and provide service logs. The customer also is required to cleanse confidential information from the logs and all of this must be done within a 30 day period of the outage.

3Tera: This is done automatically without requiring any effort by the customer.

It is important to note here that outages are defined as availability or non availability of services and not tied to application performance achievable when services are available.

C. How Does The Credit Get Applied?

AWS: The AWS credit gets applied against future credits and is not a reimbursement of lost services. As previously stated it is the customer's responsibility to provide all of the proof and do it with a 30 day period. If the customer supplies all of the documentation and Amazon approves the outage that qualifies for the below 99.95% SLA guarantee, they will then apply a 10 percent discount on the next month's bill.

3Tera: 3Tera's credit gets applied to the current month's bill. VPDC customers automatically receive SLA service credits for any calendar month where availability falls below the targeted 99.999 percent. If availability is anywhere between 99.999 percent and 99.9 percent, a 10 percent credit applies to the whole VPDC service for the entire month. If availability is lower than 99.9 percent, a 25 percent credit applies.

Based on the above, we conclude that 3Tera provides a significantly more customer friendly SLA. The recent acquisition of 3Tera by Computer Associates proves the importance of better SLA delivery to cloud users. However, the key message we convey out of this discussion is that the level of services offered by either AWS or 3Tera

is subpar compared to what is expected from IT infrastructures within the enterprise. For example, credits for outages and their implications in cloud environments are tied to number of hours of service rented in terms of server resources (CPU, memory, network bandwidth) and not tied to lost business and revenues, unlike the situation with internal enterprise data centers. The likely reason for this situation is the commodity nature of public cloud services today and the pressure to keep infrastructure costs down. In summary, significant challenges exist in the delivery of higher service levels to enable enterprises to move mission critical workloads to cloud infrastructures.

D. Application Level Performance SLAs

Application level performance in terms of minimum guaranteed job operations per second is not part of SLA offerings by either AWS or 3Tera. Monitoring functions deployed verify only availability of VMs, not specific performance delivered by them. The onus is on customers to purchase amount of CPU, memory, network and storage resources they may need to support their users. Customers must forecast possible spikes in load for their applications and purchase the maximum amount of resources. If the purchase of a certain number of VMs with set resources does not meet performance targets, the onus is on the customer to identify such performance issues and resort to purchasing additional VMs. Recently, server virtualization software vendors have acquired technologies [4] to enable application level performance monitoring on an individual VM basis but such features are not yet available as cloud computing SLAs offered by cloud service providers.

III. CLOUD INFRASTRUCTURE SLA MAXIMIZATION APPROACHES

Cloud computing infrastructures have emerged as compelling paradigms for the deployment of distributed applications and services on the Internet due in large to the maturity and wide adoption of virtualization technologies. By relying on virtualized resources, users are able to easily deploy, scale up or down their applications seamlessly across computing resources offered by one or more infrastructure providers. More importantly, virtualization enables performance isolation, whereby each application is able to acquire appropriate fractions of shared fixed-capacity resources for unencumbered use subject to binding SLAs.

The value proposition of such cloud infrastructure offerings is highly dependent on the efficient utilization of cloud resources. For the cloud service provider, this necessitates a judicious mapping of physical resources to virtualized instances that could be acquired over prescribed, fixed periods (e.g., daily or hourly). To be flexible, a provider must be able to offer a range of such instances so as to cater to a wide range of customer needs, spelled out as SLAs defined over the various resources of the instance (e.g., CPU, memory, local storage, network bandwidth). To

be able to manage a balance between SLAs and infrastructure costs, providers today limit the set of instance choices available to customers. For example, as of February 2010, Amazon EC-2 offers seven instance types: three types of standard instances, two types of high memory instances, and two types of high-CPU instances [2]. While varied, the range of instance types available to cloud customers is unlikely to match their specific application needs. As a result, customers must “over provision” by acquiring instances that are sized to support peak utilizations. More importantly, since many applications exhibit highly variable resource utilization over time (e.g., due to diurnal workload characteristics), and given the overheads associated with resizing acquired instances, customers may end up over-provisioning over extended periods of time.

Numerous research efforts have been conducted by the academia and commercial companies to determine best approaches to improving SLA in cloud infrastructures while minimizing capital and operational expenses. We cite many such relevant works in this paper.

Work done by Tickoo, Iyer, Illikkal and Newell [5] takes a look at the challenges of modeling virtual machine (VM) performance on a datacenter server. The key considerations when modeling the performance of VMs can be summarized as follows: (a) VM performance is not only dependent on its own characteristics, but also dependent on the interference caused by the other virtual machines running on the same platform with it. One needs a method to capture the effect of these interactions. (b) The above interference can affect the use of (i) shared resources (e.g. core, memory capacity) that are visible to the operating system or virtual machine monitor directly or through performance counters and (ii) shared resources (cache space, memory bandwidth, etc) that are invisible to the operating system since they are transparent resources managed by the hardware. The modeling approach needs to be aware of both visible and invisible resource interference. (c) the specifics of virtualization technology (both hardware virtualization and software virtualization) and the scheduling disciplines adopted by the virtual machine monitor could be quite different on any given platform. The modeling approach needs to take into account the virtualization technology as well as the scheduling heuristics required.

Several other research efforts look at SLA-aware virtual resource management in cloud infrastructures. Work by Van and Tran [6] looks at the downside of the flexibility brought by virtualization in terms of the added system management complexity for IT managers. Two levels of mapping must be managed: the provisioning stage is responsible for allocating resource capacity in the form of virtual machines to application. This stage is driven by performance goals associated with the business-level SLAs of the hosted applications (e.g. average response time, number of jobs completed per unit of time). Virtual machines must then be mapped to physical machines. This VM placement problem is driven by data center policies

related to resource management costs. A typical example is to lower energy consumption by minimizing the number of active physical servers. The authors separate the VM provisioning stage from the VM placement stage within the global decision layer autonomic loop and formulate both problems as Constraint Satisfaction Problems (CSP). Both problems are instances of an NP-hard knapsack problem for which a Constraint Programming approach is a good fit. The idea of Constraint Programming is to solve a problem by stating relations between variables in the form of constraints which must be satisfied by the solution. It is noteworthy that the VM packing CSP produces the VM placement vectors which are used to place VMs on PMs (Physical Machines). The solution computes the difference with the VM placement produced as a result of the previous iteration, determines which VM needs to be migrated. An optimal migration plan is produced as described in [7] to minimize the number of migration required to reach the new VM-to-PM assignment. Minimizing the cost of a reconfiguration provides a plan with few migrations and steps and a maximum degree of parallelism, thus reducing the duration and impact of a reconfiguration. The migration cost of a VM is approximated as proportional to the amount of memory allocated to the VM.

Broboff, Kochut and Beaty [8] proposes a virtual machine placement algorithm which resorts to forecasting techniques and a bin packing heuristic to allocate and place virtual machines while minimizing the number of PMs activated and providing probabilistic SLA guarantees.

Wood, Shenoy and Venkataramani [9] propose two approaches for dynamically mapping VMs on PMs: a black box approach that relies on system-level metrics only and a grey box approach that takes into account application-level metrics along with a queuing model. VM packing is performed through a heuristic which iteratively places the highest-loaded VM on the least-loaded PM.

Studies on energy efficient resource management in virtualized cloud data centers by Beloglazov and Buyya [10] show that energy savings are achieved by continuous consolidation of VMs according to current utilization of resources, virtual network topologies established between VMs and thermal state of computing nodes. In this paper the authors present a decentralized architecture of the resource management system for cloud data centers and propose the development of the following policies for continuous optimization of VM placement: (1) Optimization over multiple system resources – at each time frame VMs are reallocated according to current CPU, RAM and network bandwidth utilization. (2) Network optimization – optimization of virtual network topologies created by intercommunicating VMs. Network communication between VMs should be observed and considered in reallocation decisions in order to reduce data transfer overhead and network devices load. (3) Thermal optimization – current temperature of physical nodes is considered in reallocation decisions. The aim is to avoid “hot spots” by reducing workload of the overheated nodes

and thus decrease error-proneness and cooling system load. Simulation results show that the highest average SLA of 89% at minimum energy consumption of 1.5 KWh is achieved with about 34,231 migrations, and with about 9% SLA violations. Limiting migrations to 3,359 results in lower average SLA of 56% with energy consumption of 1.5 KWh and only 1.11% SLA violations. Higher number for migrations seems to incur the cost of higher percentage of SLA violations because of VM performance interference. Static allocation policies result in power consumptions as high as 9.15 KWh.

Finally, research work by Ishakian, Sweha et al [11] suggests more efficient utilization of instances (such as those offered by Amazon EC2 [2]) could be achieved by appropriately co-locating applications from multiple cloud customers on the same instance. The authors note that virtualization allows both co-location and performance isolation of applications by viewing such applications as independent VMs. Such VM co-location could be done in a multitude of ways: (1) it could be offered as a (distinguishing) feature by the cloud service provider. (2) It could be used in a peer-to-peer fashion to allow cloud customers to form coalitions that benefit from co-location. The paper presents VM migration strategies for co-location as a service, impact of migration on throughput and response times. The paper shows that significant cost savings per cloud customer could be realized (through avoidance of over provisioning) by supplying cloud tenants with the means to efficiently co-locate their workloads on cloud resources. The paper uses simplified criteria of grouping tenants and does not consider more complex real life scenarios such as services to be co-hosted are of a periodic, real-time nature, the identification of groupings of tenants that can be efficiently collocated would require the development of new functionalities and services.

The above cited work indicates one common trend – that any method applied to increase SLA levels for cloud users while minimizing infrastructure expenses (server, network, storage, power, space, management) for cloud service providers requires elimination of VM migration inefficiencies in terms of time to migrate VMs, cost of such migrations in terms of resource usage, and VM performance interference.

IV. VM MIGRATION EXPENSES AND THE IMPACT OF I/O PERFORMANCE

The above research work cited in this paper ([5], [6], [7], [8], [9], [10], [11]) either ignore I/O as a parameter on the mathematical models (considers CPU and memory resource allocations per VM), or limits I/O resource availability to Gigabit Ethernet. We believe this is most likely a result of lack of higher speed I/O interface support in hypervisors such as Xen used in these experiments, the overheads in hypervisors related to faster I/O processing, and the lack of consideration that performance of many data center applications are impacted directly by availability of network and storage I/O throughput and latency. In this section, we

revisit some of the above cited papers and a few additional ones to explore resource and time related expenses incurred with VM migrations and the net contribution of I/O resources deficiencies in VMs and hypervisors in those expenses.

In the research work by Ishakian, Sweha et al [11], Gigabit Ethernet connectivity is used in the experiments. Due to Xen's inability to allocate specific I/O bandwidth to VMs, Linux filtering mechanisms are used which may have consumed CPU for processing, causing VM performance interference as cited by [5]. The approach also uses concepts of VM bundles to reduce the cost of VM migrations. VM bundles are defined either to avoiding migration of VMs with larger memory footprints or to migrate the fewest number of VMs. This is an example that shows that I/O and other server resource constraints are placed up front to find optimal solutions, giving commodity server and network I/O based operational expenses higher priority than application level SLAs. Work done by Beloglazov and Buyya [10] recommend minimization of VM migrations (due to costs associated with them) to reduce power consumption, but at the cost of lower percentage of SLA fulfillment. Van and Tran's work in [6] uses CPU and memory resources only to define size of VMs in constraint satisfaction mathematical models used. Many of the research work cited measures SLA in terms of job operations per second in conjunction with VM placement and migration strategies. However, in the measurement of SLA, none of them use I/O intensive applications like data warehousing, online transaction processing, business analytics, financial services, and high performance computing applications. Use of such applications can highlight the need for efficient I/O to serve applications that run in VMs and the need to eliminate performance interference between VMs and VM migrations in such use cases [5].

The paper Experimental Study of Virtual Machine Migration in Support of Reservation of Cluster Resources by Zhao and Figueiredo [12], seeks to provide a model that can characterize the VM migration process and predict its performance, based on a comprehensive experimental analysis. The results show that, given a certain VM's migration time, it is feasible to predict the time for a VM with other configurations, as well as the time for migrating a number of VMs. The paper also shows that migration of VMs in parallel results in shorter aggregate migration times, but with higher per-VM migration latencies. In their experiments, for VMs with 512MB memory sizes, the time needed for migrating a single VM is 8.5 seconds, while the time needed per VM when 4 VMs were migrated in sequence is 11.5 seconds.

Figure 1 below shows the migration time and performance degradation when four VMs were migrated in sequence, each with 512MB memory and a CPU-intensive benchmark running inside.

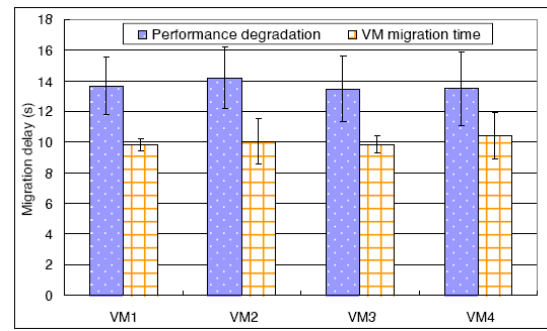


Fig. 1. The migration time and performance degradation when four VMs were migrated in sequence, each with 512MB memory and a CPU-intensive benchmark running inside (courtesy [12]).

Figure 2 shows the migration time and performance degradation when four VMs were migrated in sequence, each with 512MB memory and a memory-intensive benchmark running inside.

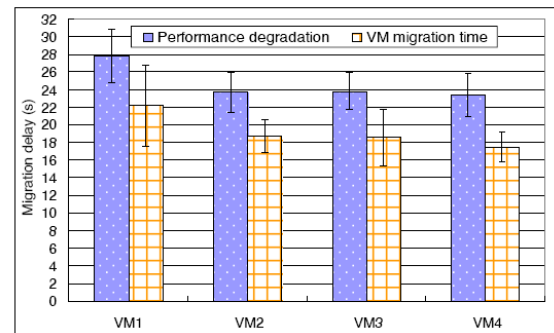


Fig. 2. The migration time and performance degradation when four VMs were migrated in sequence, each with 512MB memory and a memory-intensive benchmark running inside (courtesy [12]).

Experiment results also show that parallel migration is faster, and the advantage becomes larger when more VMs are migrated together. For VMs with 256MB of memory, the speed up is 1.4 times for 4 VMs, and 1.6 times for 8 VMs. For VMs with 512MB memory size, the speedup is 1.3 times for 4 VMs, which is less than that of the smaller VMs. This is because the advantage from parallel migration is mostly from overlapping the suspend and resume phases of multiple VMs, since the copy phase is bounded by the available network bandwidth (Gigabit Ethernet). For larger VMs, their migrations are dominated by the copy phase and thus cannot gain much from the parallel migration because of network I/O constraints.

Some inferences from this paper are: (1) migration time increases significantly with increasing memory sizes per VM and the number of VMs being migrated in sequence; (2) memory-intensive tasks incur higher performance degradation and consume more VM migration time; (3) most VM migration time incurred is in the copy phase of a VM migration (suspend and resume are the other phases); and (4) parallel migration of VMs can be significantly faster

if network I/O bottlenecks can be eliminated.

Next we look at work done by Huang, Gao, Liu et al [13] on the impact of I/O on VM migration time. Currently, most VM environments use the Socket interface and the TCP/IP protocol to transfer VM migration traffic. In this paper, the authors propose a high performance VM migration design by using RDMA (Remote Direct Memory Access). RDMA is a feature provided by many modern high speed interconnects (such as Ethernet and InfiniBand) that are currently being widely deployed in data-centers and clusters. By taking advantage of the low software overhead and the one-sided nature of RDMA, the proposed design significantly improves the efficiency of VM migration. The evaluations using a prototype implementation over Xen show that RDMA can drastically reduce the migration overhead: up to 80% on total migration time and up to 77% on application observed downtime. We devote the next section in highlighting excerpts of this breakthrough work that promises to solve the VM migration cost challenges highlighted in this paper and make application of VM migration technologies cost effective to deliver significantly higher SLAs while minimizing cloud infrastructure expenses.

V. VM MIGRATION EFFICIENCIES OVER HIGH PERFORMANCE, LOW LATENCY RDMA NETWORKS

High speed interconnects, such as InfiniBand [14], iWARP for 10 GigE [15] and RoCE for 10 GigE [16], open up an opportunity for significant improvements in the efficiency of VM migration.

The RDMA technology offered by these interconnects is an ideal match for the task at hand. RDMA allows direct data placement of data from one node's memory space into another. This is attained without memory copies on the local side and with no involvement of the remote CPU. The above, combined with low latency implementations over high throughput links (10/40Gig) makes it particularly suitable to the scenario discussed where a vast amount of memory needs to be moved across nodes. This application of RDMA allows for very low migration latencies with minimal consumption of compute resources devoted to the migration task itself.

Access to the RDMA interfaces is implemented through a highly optimized SW APIs that introduces no performance penalties to the communication. Migration is typically controlled by the hypervisor which gets its own dedicated access to the IO device with no data copies or context switches involved. This allows migration to be carried out with minimum impact on guest operating systems and hosted applications. As a side note, even for regular VM IO purposes, virtualization is natively built into these RDMA interfaces which abstract IO access through channels that can be exposed directly to the VMs.

The use of RDMA efficiently reduces the time required to transfer the VM memory pages and this leads to immediate savings on total VM migration time. The paper by Huang, Gao, Liu et al [13] studies RDMA based VM migration. The authors analyze the challenges to achieve efficient VM

migration over RDMA, including protocol design, memory registration, non-contiguous data transfer, network QoS, etc. Evaluations with our prototype implementation of Xen migration over OpenFabrics software [17] and RDMA based protocols are able to significantly improve the migration efficiency. The OpenFabrics software distribution includes support for sockets and RDMA interfaces over InfiniBand, RoCE and iWARP. The experiments in this paper use InfiniBand as the data link layer and interconnect over which VM migrations are conducted. For example, compared with the original Xen migration over TCP/IP sockets (using IPoIB [18]), the proposed design over InfiniBand RDMA reduces the impact of migration on SPEC CINT 2000 Benchmarks [19] by an average of 54% when the server is lightly loaded, and an average of 70% when it is heavily loaded.

The experiments compare total migration time achieved over sockets (IPoIB), RDMA read and RDMA write operations. Figure 3 shows the total migration time needed to migrate a VM with varied memory configurations.

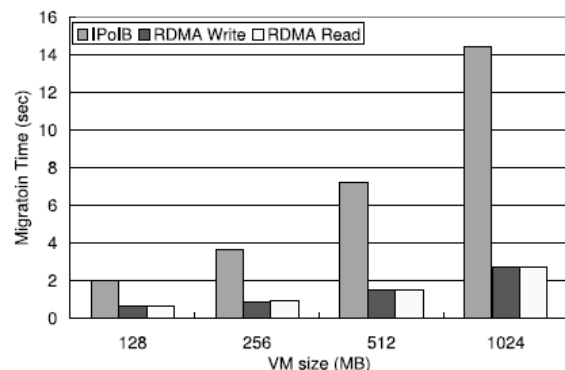


Fig. 3. Total Migration Time (courtesy [13]).

As we can see, due to the increased bandwidth provided by InfiniBand and RDMA, the total migration time can be reduced by up to 80% by using RDMA operations.

Further experiments on CPU utilization show that compared to VM migration over sockets (IPoIB), RDMA-based migration reduces the impact on applications by up to 89%, and an average of 70%. Freeing up CPU resources can positively impact VM performance (through reduction of performance interference as highlighted in [5]). Finally, the paper highlights that RDMA based migration can significantly reduce the migration overhead observed by applications hosted on both the migrating VM and the non-migrating VMs. This is especially true when the server is highly loaded and has less CPU resources to handle the migration traffic, and less interference with non-migration related I/O traffic is required to ensure application performance.

VI. CONCLUSIONS AND FUTURE WORK

Cloud infrastructures have the challenge of delivering higher levels of SLAs to cloud users while minimizing capital and operational expenses with cloud infrastructures incurred by cloud service providers. The two goals are divergent with the latter aiming to use minimum number of servers through

tighter VM packing in physical servers while delivering higher application level SLAs require optimum placement of VMs with maximum available resources per VM in physical servers. This is evident today in cloud services offered by major cloud service providers where SLA seems to be compromised in favor of lower cost of operations. Several research topics and papers are discussed in this paper. These experiments utilize constraint based mathematical models that try to maximize SLAs while minimizing physical server usage. Such experiments use VM placement and VM packing algorithms which result in use of VM migration strategies to reach the goals. We evaluated the cost of various VM migration strategies in terms of migration times, performance degradation, impact on power efficiencies, success in meeting SLAs, number of VM migrations needed, impact of moving CPU-intensive and memory-intensive VMs and impact of migrating VMs in sequence versus parallel. We infer that certain VM migration strategies such as migrating VMs in parallel can be efficient if I/O bottlenecks can be eliminated. Next, we looked at research work applying high performance and low latency RDMA-based I/O technologies for VM migration and the results show compelling improvements in VM migration efficiencies in terms of both VM migration time and reducing interference to performance of VMs and applications running in the VMs. Based on these findings, we conclude that use of RDMA technologies for VM migrations can significantly benefit both the cloud service providers in terms of more efficient use of their cloud infrastructure while allowing cloud users to enjoy higher levels of SLA.

The compelling VM migration performance and efficiency results in [13] were obtained using the OpenFabrics RDMA verbs interface which is also available over 10 Gigabit Ethernet technologies. RoCE (RDMA over Converged Ethernet) implements the same RDMA interface and transport services as used in the above experiments using InfiniBand and as such RoCE based 10 Gigabit Ethernet technologies may be used to achieve similar VM migration efficiency results.

Use of SRIOV (single root I/O virtualization) technologies [20] with high performance I/O technologies can enable higher application level performance (jobs per second, response times) when they are running in VMs while reducing hypervisor overheads and performance interference [5]. In the cited works in this paper that measure SLA at the application level, significant performance SLA benefits can be achieved while minimizing interference caused by VM migrations. For example, a 10 Gigabit per second I/O pipe available in Ethernet or a 40 Gigabit per second I/O pipe available in InfiniBand can be divided into smaller and isolated I/O pipes with maximum allowable throughput and minimum latency parameters and such smaller pipes can be dedicated to VMs running applications and VM migration functions. The smaller I/O pipes to VMs are isolated from each other preventing performance interference in one VM because of load spikes in another. Also, if the amount of bandwidth and latency available to VMs and VM migrations can be dynamically adjusted based on transient SLA, user and

application load scenarios as prevalent in cloud infrastructures, delivery of even higher application performance level SLAs can be possible while minimizing capital and operational expenses – a win-win for both cloud users and cloud service providers.

REFERENCES

- [1] Wikipedia on Service Level Agreement, SLA: http://en.wikipedia.org/wiki/Service_Level_Agreement
- [2] <http://aws.amazon.com/ec2-sla/>
- [3] <http://www.3tera.com/News/Press-Releases/Recent/3Tera-Introduces-the-First-Five-Nines-Cloud-Computing.php>
- [4] Purchase of B-hive by VMware: http://www.vmware.com/company/news/releases/bi_hive.html
- [5] Modeling Virtual Machine Performance: Challenges and Approaches by Tickoo, Iyer, Illikkal and Newell, Intel Corporation
- [6] SLA-aware virtual resource management for cloud infrastructures, Hien Nguyen Van, Fr'ed'eric Dang Tran, Orange Labs
- [7] F. Hermenier, X. Lorca, J.-M. Menaud, G. Muller and J. Lawall. Entropy: a Consolidation Manager for Cluster. In proc. of the 2009 International Conference on Virtual Execution Environments (VEE'09), Mar. 2009.
- [8] N. Bobroff, A. Kochut and K. Beaty. Dynamic Placement of Virtual Machines for Managing SLA Violations. 10th IFIP/IEEE International Symposium on Integrated Network Management, May 2007.
- [9] T. Wood, P. Shenoy, A. Venkataramani, and M. Yousif. Black-box and Gray-box Strategies for Virtual Machine Migration. 4th USENIX Symposium on Networked Systems Design and Implementation, 2007.
- [10] Energy Efficient Resource Management in Virtualized Cloud Data Centers, Anton Beloglazov and Rajkumar Buyya, Cloud Computing and Distributed Systems (CLOUDS) Laboratory, Department of Computer Science and Software Engineering, The University of Melbourne, Australia
- [11] Co-location as a Service: Strategic and Operational Services for Cloud Co-location, VATCHE ISHAKIAN, RAYMOND SWEHA, JORGE LONDO'NO, AZER BESTAVROS, Computer Science Dept, Boston University, USA
- [12] Experimental Study of Virtual Machine Migration in Support of Reservation of Cluster Resources, Ming Zhao Renato J. Figueiredo, Advanced Computing and Information Systems Laboratory (ACIS) Electrical and Computer Engineering, University of Florida
- [13] High Performance Virtual Machine Migration with RDMA over Modern Interconnects, Wei Huang, Qi Gao, Jiuxing Liu, Dhableswar K. Panda, Computer Science and Engineering, The Ohio State University and IBM T. J. Watson Research Center
- [14] InfiniBand Trade Association, www.IBTA.org, InfiniBand Architecture Specification, Release 1.2.
- [15] iWARP, Internet Wide Area RDMA Protocol, <http://en.wikipedia.org/wiki/IWARP>
- [16] RoCE – RDMA over Converged Ethernet, InfiniBand Trade Association, http://www.infinibandta.org/content/pages.php?pg=press_room_item&ec_id=663
- [17] OpenFabrics RDMA Protocols through OFED software, www.openfabric.org
- [18] IETF IpoIB Workgroup. <http://www.ietf.org/html.charters/ipoib-charter.html>.
- [19] SPEC CPU 2000 Benchmark. <http://www.spec.org/>.
- [20] SRIOV specification, http://www.pcisig.com/specifications/iov/single_root/

Sujal Das is Sr. Director, Product Management, at Mellanox Technologies. Email: sujal@mellanox.com.

Michael Kagan is Chief Technology Officer at Mellanox Technologies. Email: michaelk@mellanox.com.

Diego Crupnicoff is Senior Architect at Mellanox Technologies. Email: diego@mellanox.com.