論文の内容の要旨

Efficient Replication Mechanisms for Highly Available Virtual Machines

(仮想マシンの可用性向上のための効率的なレプリケーション機構)

氏名　　バリ　ゲローフィ

With the recent increase of cloud computing's prevalence, the number of online services deployed over virtualized infrastructures has experienced a tremendous growth. At the same time, however, the latest hardware trend of growing component number in current computing systems (e.g., data-centers, high-end computer clusters, etc.) renders hardware failures common place rather than exceptional.Hardware failures can lead to a degradation in performance to end-users due to service unavailability and can result in losses to the business, both in immediate revenue as well as in longterm reputation. Thus, software solutions capable of masking hardware failures in a transparent manner are becoming more and more important. Replication at the Virtual Machine (VM) layer is an attractive technique towards accommodating VM installations with high availability, primarily, because it provides seamless failover for the entire software stack executed inside the virtual machine, regardless the application or the underlying Operating System (OS).

One particular approach, checkpoint-recovery based VM replication has received a lot of attention during the last few years. Checkpoint-recovery based replication of virtual machines is attained by capturing the entire execution state of the running VM at relatively high frequency in order to propagate changes to the backup machine almost instantly. This solution, essentially, keeps the backup machine nearly up-to-date with the latest execution state of the primary machine so that the backup can take over the execution in case the primary fails. While checkpoint-recovery based replication is inherently capable of tackling with symmetric multiprocessing (SMP) VMs, i.e., virtual machines with multiple virtual CPUs (vCPUs), due to the large amount of state that needs to be synchronized between the primary and the backup machines, performance degradation of the applications executed in replicated VMs can be significant even on uni-processor setups.

This dissertation makes four main contributions towards providing good performance and ensuring high availability at the same time. As one of the major factors of replication overhead is the large amount of state that needs to be transferred over the network, efficient data compression can potentially mitigate overhead. First, this thesis presents a novel approach for decreasing network traffic during synchronization between the primary and backup machines by a lightweight compression method that exploits data self-similarity inside

the virtual machine's memory image. A bit-projection based hash function is proposed, upon which, a content based hash table is built for the purpose of finding memory areas that are similar to the dirtied memory. The similar memory areas are then utilized to obtain and transfer only differences, instead of entire memory pages.

Second, the potential advantages of utilizing features of high-performance interconnects, such as Remote Direct Memory Access (RDMA) and OS-bypass communication, are explored in the context of VM replication.

Naturally, different workloads exhibit different behavior in terms of memory usage and I/O patterns. When exactly checkpoints are taken has substantial impact on the efficiency of the replicated virtual machine. As opposed to previous studies, which employ fixed checkpoint frequency, this thesis investigates the impact of dynamic checkpoint scheduling. Our third contribution, workload adaptive checkpoint scheduling, initiates checkpoints with respect to the memory and I/O behavior of the given workload, as well as to the network bandwidth available for replication.

Finally, TCP transmission performance of replication virtual machines is considered. We revisit the basic replication protocol and extend it with speculative communication. Speculative communication fabricates and delivers speculative acknowledgments to the VM in response to its TCP packets so that the guest's TCP stack can progress with TCP transmission. Furthermore, we propose replication aware congestion control, an extension to the guest OS' TCP stack that aggressively fills up the replication buffer of the VMM so that a large amount of speculative packets can be sent to the backup host and released earlier to the clients.

These improvements, which have been implemented in the Linux Kernel Virtual Machine (KVM), are evaluated under a variety of workloads that show superior performance compared to existing replication solutions.