Performance improvements in a large scale virtualization system

The Worker Nodes on Demand Service (WNoDeS), developed by INFN, is a framework designed to offer local, grid or cloud-based access to computing and storage resources, preserving maximum compatibility with existing computing center policies and workflows.WNoDeS has been running in production at the INFN Tier-1 located at CNAF since November 2009, where it currently manages more than 2,000 dynamically created Virtual Machines; WNoDeS is also being deployed at several other Italian sites. WNoDeS makes extensive use of virtualization technologies to offer resource polymorphism. Based on more than a year of experience in running thousands of VMs in a production environment used by several international collaborations, this work shows the optimizations we have been investigating and implementing at the virtualization layer. These optimizations increase the adaptability of WNoDeS to demanding applications like those run by High-Energy Physics experiments. Description of the work: WNoDeS is a layered framework re-using as many supported and production solutions as possible. The WNoDeS virtualization layer is managed by Linux KVM; KVM was chosen after a careful evaluation of competing virtualization solutions, including an extensive suite of performance tests, was performed. On the one hand, the WNoDeS framework is designed to manage thousands of dynamically instantiated Virtual Machines; on the other hand, with current multi-core CPUs, having 10 or 20 running VMs on a single hardware platform is a concrete possibility. Performance and scalability issues that have to be considered at the virtualization layer are thus centered across two main areas: 1) performance of individual VMs. Each and every VM must be capable of addressing the CPU and I/O demands of high-throughput applications, such as those currently run by LHC experiments. 2) handling of VM images and interaction of VMs with shared storage file systems. This item has two sub-points: 2.1) WNoDeS VMs may need to access and work on data residing on shared storage, and they need to do so in an efficient and scalable way; 2.2) The WNoDeS VMs themselves are stored - in read-only mode - on shared storage. They are downloaded to and cached on WNoDeS hypervisors when needed. It is important that this download happens as efficiently as possible, so that the shared file system is not overloaded, nor is network bandwidth saturated. This talk will explain how we are addressing the two points above. In particular, for the first point we will report our testing and deployment experience on several improvements that were recently introduced to KVM, such as KSM, transparent hugepages, virtio, SR-IOV (for network PCI-passthrough) and VMs on LVM partitions. For the second point, we will first describe how each WNoDeS VM was originally directly interfaced to a parallel distributed file system like GPFS, showing the merits and the issues of such an approach; we will then examine how alternative solutions to interface VMs to a shared storage can be chosen and will show our experience with them. These solutions include the adoption of read-only filesystems with an http back-end like CVMFS, of other FUSE-based filesystems like sshfs, or of NFS/GPFS interfaces. The overall goal of these solutions is to limit the growth of shared parallel file system clusters and at the same time to ensure manageability, reliability and scalability of a large VM-based cluster.

Primary authors : SALOMONI, Davide (INFN CNAF) ; CHIERICI, Andrea (INFN CNAF)