Snowflock: Virtual Cluster Technology for Bioinformatics Applications

Virtual machine (VM) technology has become a popular method for simplifying management and sharing of physical computing resources. Platforms such as VMWare and Xen allow multiple VMs with different operating systems and configurations to co-exist on the same physical host in mutual isolation. VMs have additional advantages within bioinformatics: many bioinformatics applications use databases that are much larger than the program itself, and VM managers allow a running VM to migrate to a host that is close to the database, thereby speeding up the computation. One limitation of VMs is that a single virtual machine can only use processors located on a single physical machine, and cannot take advantage of compute clusters. This drawback is especially apparent in bioinformatics, where many tasks are parallelizable. To overcome this limitation we have developed Snowflock, a virtual machine manager tailored to bioinformatics applications.

Snowflock implements a novel cloning mechanism for VMs, allowing users to create identical, networked copies of a running VM on demand. These copies form a virtual cluster, distributed dynamically across physical hosts and connected by a secure virtual network. The virtual network allows the VMs of each virtual cluster to communicate securely in isolation from the physical host network, and to access shared resources such as large biological databases.

The cloning mechanism is implemented to provide fork-like semantics, allowing virtual clusters to be created, used, and discarded much as processes are created, used and discarded in a typical Unix environment. Cloning is highly optimized: exploiting hardware multicast and intelligent page/block fetching allows a hundred or more clones to be up and running at full speed in a few seconds. The scheduling of VMs on a physical cluster can be handled by any job control system, allowing integration with existing workflows on the physical hosts. An additional capability of Snowflock is the migration of a virtual machine across the internet, allowing users without access to a local compute cluster to bring their programs, scripts and environment to a remote site with the required data and computational power in a secure manner. This greatly simplifies the sharing of hardware clusters for custom bioinformatics scripts and applications.