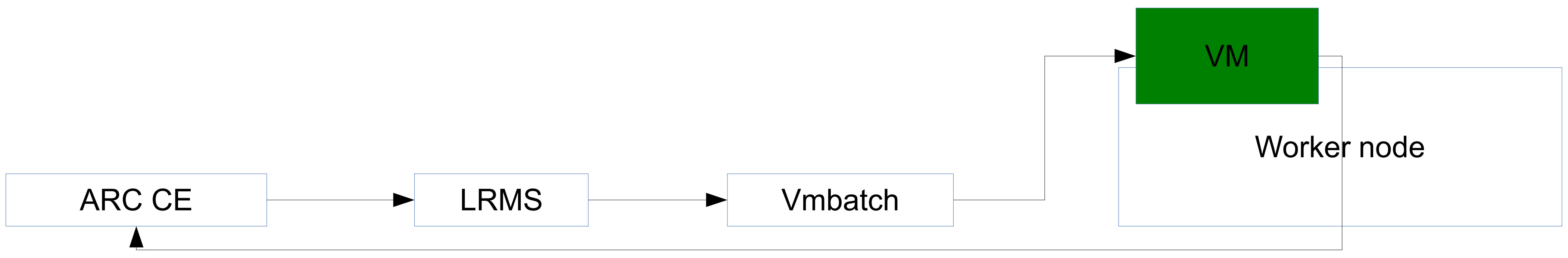
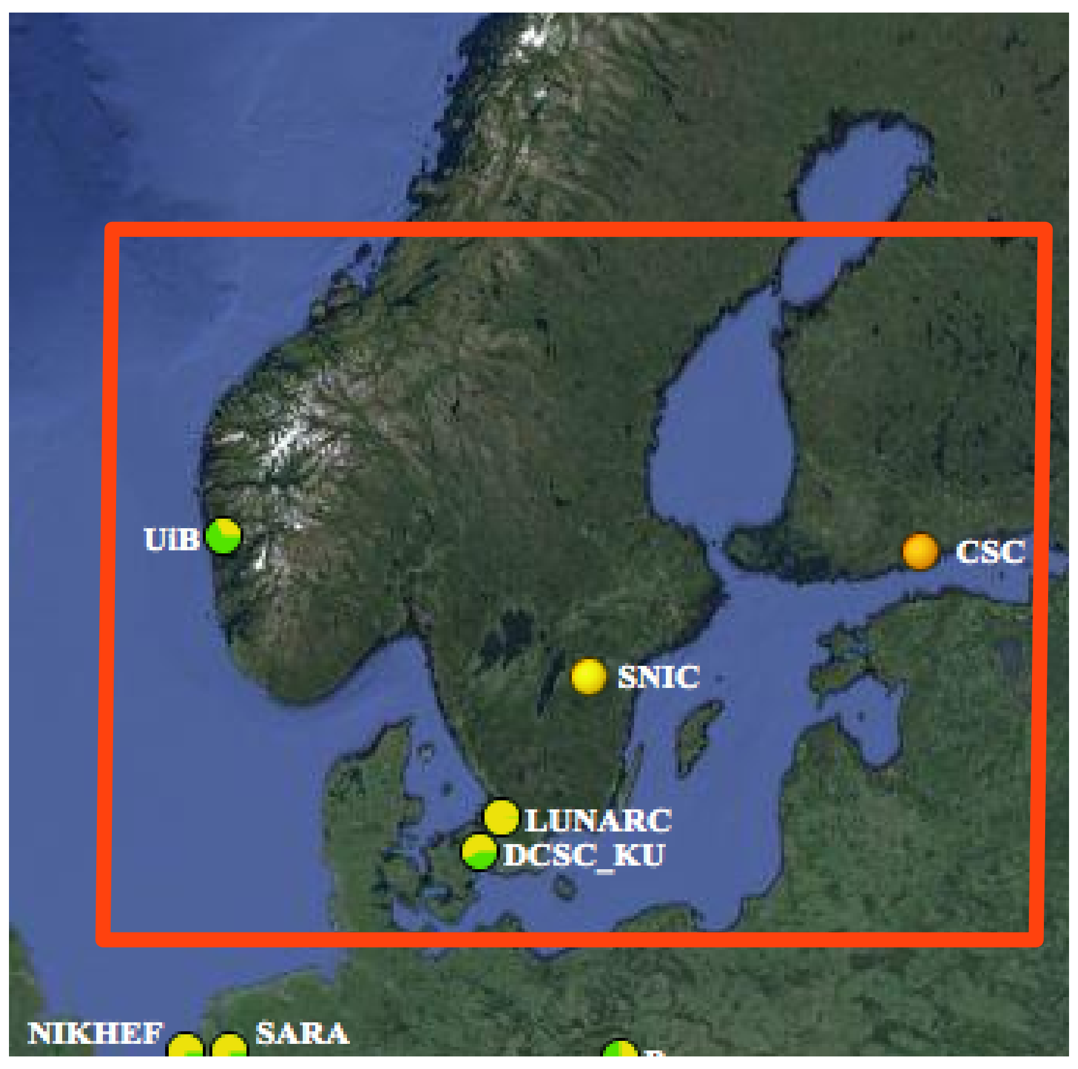
A new provisioning method to explore different virtualisation methods for ALICE grid jobs over ARC

B. Wagner[1], B. Kileng[2], for the ALICE Collaboration
[1]Universitetet i Bergen, [2]Høgskolen i Bergen

The distributed Nordic Tier-1 for LHC uses ARC[1] as internal computing grid middleware. ALICE[2] uses its own grid middleware AliEn[3] to distribute jobs and the necessary software application stack. We are developing a testbed and possible framework to bridge those different middleware systems with virtualisation.





Nordic Tier-1

The Nordic countries decided to bundle computing centers from different countries and form a distributed Tier-1 centre. It is steered by the Nordic data grid facility (NDGF[4]) part of NeIC[5].

CernVM

CernVM[6] is a project to provide a software environment for all CERN experiments. It consists of a small core virtual machine and the CernVM file system (cernvmfs). It is a read-only network filesystem based on http. Implemented as a FUSE[7] module, it appears as if the experiment specific software stack is available locally, but only when a file is accessed it gets fetched and cached.

CernVM supports several methods of contextualisation. Vmbatch uses the cdrom method by generating a small custom ISO image and attaching it to the VM via additional cdrom tags in the domain-xml configuration file.

References
[1] ARC - M. Ellert et al., Future Generation Computer Systems 23 (2007) 219 2013240
[2] ALICE - K. Aamodt et al., 2008 JINST {\bf 3} S08002
[3] AliEn – S. Bagnasco et al., J. Phys.: Conf. Ser. {\bf 119} (2008) 062012
[4] NDGF - http://www.ndgf.org
[5] NeIC - http://www.nordforsk.org/no/programs/nordic-einfrastructure-collaboration-neic
[6] CernVM - P. Buncic et al, 2010 J. Phys.: Conf. Ser. 219 042003
[7] fuse - http://fuse.sourceforge.net
[8] Vmbatch - Subversion repository http://eple.hib.no/svn/vmbatch/tags/latest/
[9] ViBatch - https://ekptrac.physik.uni-karlsruhe.de/trac/BatchVirt/wiki/ViBatch
[10] Xen - P. Barham, et al. Xen and the Art of Virtualization,
19th ACM Symposium on Operating Systems Principles, 2003
[11] libvirt - http://libvirt.org/index.html
[12] Torque - G. Staples SC '06
Proceedings of the 2006 ACM/IEEE conf. on Supercomputing
[13] qcow - http://www.linux-kvm.org/page/Main_Page

Vmbatch

A testbed system has been developed (source available see Ref. [8]), based on the ViBatch[9] system, which creates virtual machines as worker nodes in a batch system.

Most tests were done with Xen[10], but the tool is based on the libvirt library[11] and supports also other hypervisors.

We use TORQUE as the cluster resource management system[12] and use its prologue and epilogue scripts to start and stop the guest.

A shared NFS file system is used for data exchange between a host and its guest systems.

In the test environment a DHCP server is used for easier guest network connectivity, but it also works with a static network set up.

There are no site specific configuration to be done to the guest VM, but there are requirements concerning the guest disk layout.

A job from a dedicated queue will cause Vmbatch to start a virtual machine on a worker node running the job. The job will be run inside this newly created guest system by using an SSH connection into the guest.

At job start a TORQUE prologue script will create new disk images based on template images. We use Qemu Copy On Write disk images (qcow[13]) created by the Xen tools.

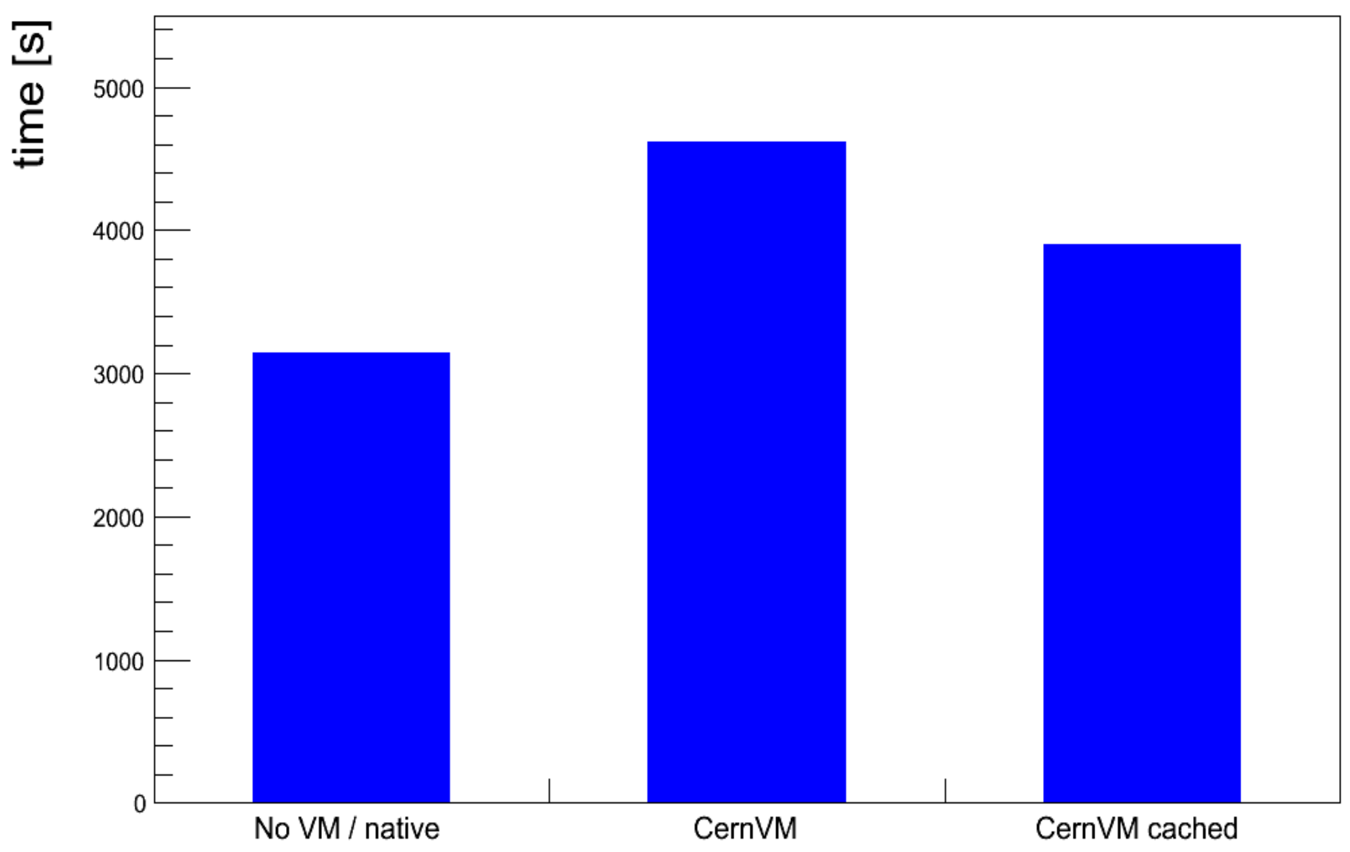
New qcow use an existing qcow image as backing image. This is a very fast method for copying large disk images. All disk writes will go the the new image but unmodified reads will come from the backing image.

The script will also set up Vmbatch specific system services in the VM and configure its network.

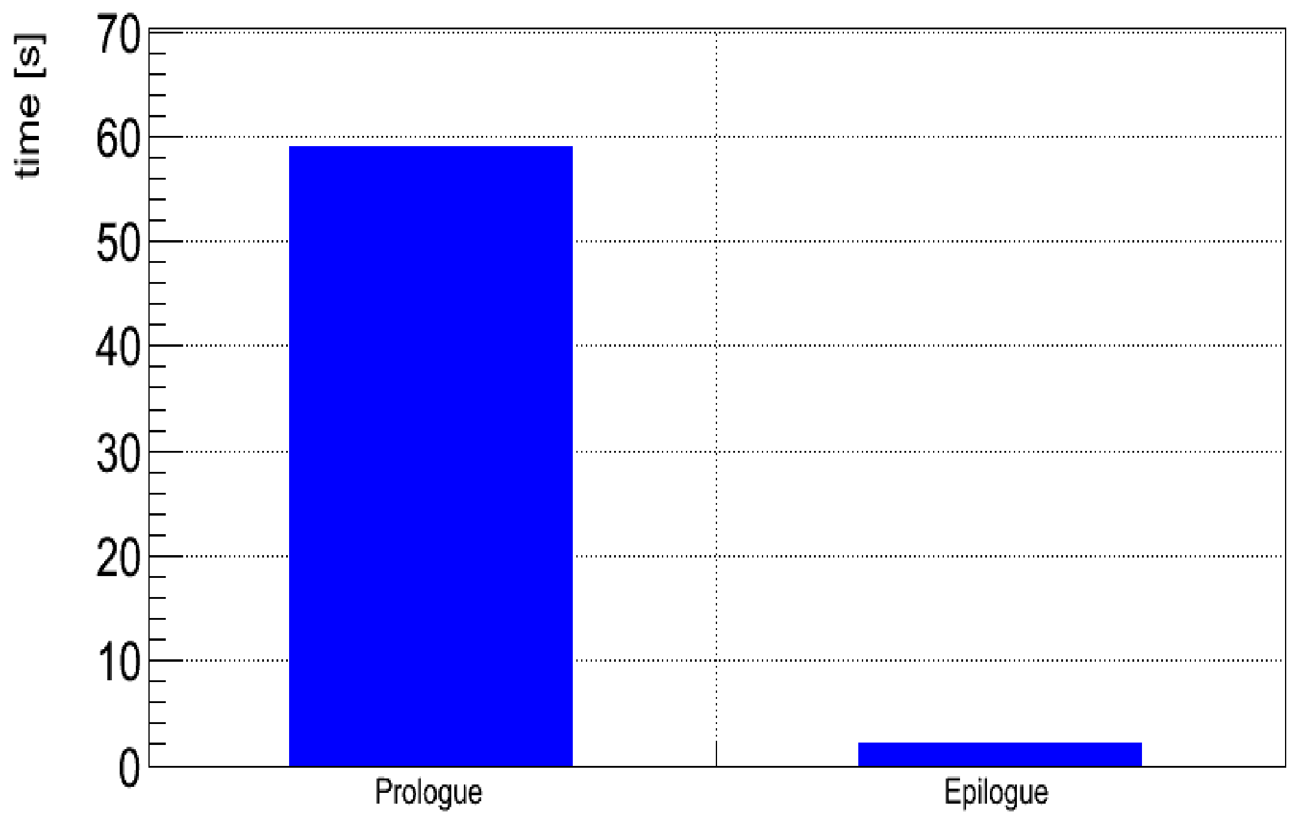
During the VM startup the Vmbatch specific system services will create the submitting user, mount the NFS share and create a specific file on the NFS share as a signal to the prologue script that the guest has started. When the guest is up and running, the prologue script will exit.

When the job is finished a TORQUE epilogue script will destroy the guest, delete the disk images and clean up lock files.

Run Time (averaged)



Prologue / Epilogue Times



Results

As a first test job we chose PbPbbench from the ALICE software analysis framework. It simulates and reconstructs lead-lead collisions.

Different configurations were measured and show that the setup time of the virtual machine in comparison to the run time of the job is negligible. Notice the different time scales in the Figure

For CernVM one has to distinguish between the first run time of the VM which has to fetch the needed files from remote storage and the consecutive jobs that will fetch most of the files from the caches.

Conclusion

The first results show that the Vmbatch system is a lightweight tool to use dynamic virtualisation in a batch processing setup.

The default settings reflect common compute cluster setups, so it works often without big configuration changes.

The fast copy on write mechanism is key to minimise the setup phase.

It is easily extendable to different underlying virtualisation methods.