The HEPiX Virtualisation Working Group: Towards a Grid of Clouds

Tony Cass
On behalf of the HEPiX Virtualisation Working Group
CERN, CH-1211 Geneva 23, Switzerland
E-mail: tony.cass@cern.ch

Abstract. The use of virtual machine images, as for example with Cloud services such as Amazon’s Elastic Compute Cloud, is attractive for users as they have a guaranteed execution environment, something that cannot today be provided across sites participating in computing grids such as the Worldwide LHC Computing Grid. However, Grid sites often operate within computer security frameworks which preclude the use of remotely generated images. The HEPiX Virtualisation Working Group was setup with the objective to enable use of remotely generated virtual machine images at Grid sites and, to this end, has introduced the idea of trusted virtual machine images which are guaranteed to be secure and configurable by sites such that security policy commitments can be met. This paper describes the requirements and details of these trusted virtual machine images and presents a model for their use to facilitate the integration of Grid- and Cloud-based computing environments for High Energy Physics.

1. Introduction

The introduction by Amazon of their Simple Storage Service (S3) and Elastic Compute Cloud (EC2) service in 2006 was noted with interest by many in the High Energy Physics community and proposals to facilitate integration of HEP computing with Amazon services were being presented at CHEP in 2007 [1] with actual implementations being presented in a CHEP plenary session in 2009 [2]. These successful examples soon led to requests for the introduction of a similar facility for the instantiation of user-prepared virtual machine images at dedicated computing centres, notably CERN and the other sites participating in the Worldwide LHC Computing Grid.

Many of these computing sites, however, resisted such requests, at least initially, mainly for two reasons related to computer security:

- resources allocated to physics computing could not be isolated from other site resources and thus remotely-generated images, particularly with the potential for uncontrolled root privileges, posed a threat to these other resources; and
- participation in computing Grids such as the WLCG required sites to maintain detailed logs [3] in order to facilitate tracing and analysis of any security incidents—a requirement that could not be met for work that was carried out in remotely generated images.

1 w3.hepix.org/virtualization
Following discussions of these issues at meetings of the HEPiX community of HEP computing institutes [4], notably in Spring [5] and Fall [6] 2009, a dedicated Virtualisation Working Group [7] was established to make proposals for solutions that would enable sites to instantiate of remotely generated virtual machine images without compromising site security and with the ability to respect their obligations under Grid security policies. This has been done with proposals established in the areas of virtual image generation, contextualisation and exchange. After a brief explanation of the proposals in the different areas, a model is presented for the use of trusted machine images to facilitate the integration of Grid- and Cloud-based computing environments for High Energy Physics.

2. Proposals of the HEPiX Virtualisation Working Group

2.1. Image Generation and Endorsement
A key point to note for High Energy Physics is that there is no real need for the many hundreds, if not thousands, of individual physicists associated with a particular experiment to generate individual virtual machine images, but rather for a small number of people to generate a tailored image for that experiment which can then be used by others. In this way experiments can be sure that analysis, reconstruction or other tasks run in a uniform environment. Given this point, the HEPiX Virtualisation Working Group introduced the concept of image endorsers, individuals who could be trusted by sites to generate images that would respect their requirements in terms of computer security. Specifically, individuals endorsing a virtual machine image would guarantee that their images would be

- free of any known security flaws—and updated rapidly should new relevant flaws be identified,
- free of any accounts or credentials that would provide “back-door” access to an image after instantiation, and
- configurable, at contextualisation time, such that sites are able to maintain the detailed log required by Grid security policies.

These proposals, with adaptations to cater for the use case where a third-party is instantiating images on behalf of a Grid Site, became the Joint Security Policy Group’s policy on Trusted Virtual Machines [8].

2.2. Image Contextualisation
As noted above, sites must be able to configure images generated elsewhere to be compatible with the site infrastructure for recording, for example, syslog information such that traceability requirements are met for Grid jobs running in such images. This ability is provided in trusted images by providing hooks for site contextualisation at the time an image is instantiated. Two options are proposed for the contextualisation hooks.

1. Trusted images should attempt to mount a CDROM image provided by the site and then invoke site-provided scripts in this image before and then after network initialisation. Similarly, trusted images will also attempt to run site-provided scripts on the CDROM image at shutdown time.

2. For compatibility with Amazon’s EC2, trusted images should also attempt to download configuration text data from the standard non-routable IP address and use these, via a HEPiX provided plugin for the amiconfig tool [9].

For those interested, the precise details of the required contextualisation steps—which may be updated for compatibility with future cloud computing environments, are documented at HEPiX VWG site.
2.3. Image Cataloguing and Distribution

Ensuring a secure and trustworthy transfer of images from an endorser to sites is clearly important: sites must be sure the images they are instantiating locally are indeed images that have been endorsed by someone that they trust. The HEPiX Virtualisation Working Group is essentially as follows.

1. Image endorsers maintain a Virtual Machine Image List which contains metadata concerning all images they currently endorse. In addition to the necessary operational information (such as image size, operating system and tag), the metadata for an image includes a secure hash (sha512) of the image and a uri to enable image download. Additionally, the image list includes a tag identifying the limit of the validity for the image list itself—i.e. image endorsers are expected to regularly and actively confirm that images they have endorsed are still valid.

2. Virtual Machine Image Lists are made available on an authenticated host which sites consult periodically.

3. New or changed images are then downloaded by sites as necessary, stored and catalogued in a local repository and made available for instantiation.

Again, a more precise explanation of this process is available from HEPiX VWG for those interested in implementing a compatible service.

Owen Synge, of the Deutsches Elektronen-Synchrotron (DESY) in Hamburg, has delivered a working reference implementation of the HEPiX procedure, hepixvimisubscriber, available at [10]. In addition to automating the procedure of image list creation for image endorsers, this tool allows sites to monitor the image lists of endorsers they trust and automatically download any new or updated images. It is important to realise that such an automation is possible with the HEPiX proposal: since the concept of image endorsement implies a trust relationship between sites and endorsers, there is no need for a heavy verification process each time an image is produced or updated.

Michel Jouvin of the Laboratoire de l’Accélérateur Linéaire (LAL) in Paris has provided an integration of the hepixvimisubscriber tool with the StratusLab Marketplace [11], automatically making available any newly provided images for use with StratusLab interfaces. A similar integration of hepixvimisubscriber with OpenStack’s Glance component [12] is currently underway at CERN.

3. A possible scenario for the use of Trusted Virtual Images

An exciting possibility should the concept of trusted virtual machine images permit the use of remotely generated images at a large number of WLCG sites is for the distributed image to directly connect to the pilot job frameworks of the LHC experiments—e.g. DIRAC for LHCb or PANDA for ATLAS. Experiments can exploit CernVM [13] to generate such images rather easily with the CernVM File System [14] being used to ensure that images do not need to be refreshed frequently to follow the updates of the experiment software frameworks.

An illustration of how this might work is shown in Figure 1. Instead of maintain complicated batch scheduling systems as now, sites instantiate copies of VO provided virtual machine images as necessary to ensure that computing resources are shared out amongst the VOs they support according to the relevant commitments. Once instantiated, these virtual machines form a dynamically changing set of resources available to experiments on which they can, using their pilot job frameworks, schedule work from their members in a way that meets their priorities. Such an implementation would cleanly separate the different functions and responsibilities of Grid sites and experiments (or virtual organisations) and combine the best aspects of the Grid and Cloud computing models.
4. Conclusion

By introducing the concept of an image endorser, a person who undertakes to guarantee to sites participating in a computing Grid that images they provide are both secure and configurable such that sites can ensure compliance with relevant computer security policies, the HEPiX Virtualisation Working Group has proposed a practical way in which virtual machine images can be used across computing Grids to ensure a computing environment that, whilst not tailored to individual users, is guaranteed to be common for a given set of users—such as members of a given experimental collaboration.

This work, together with CernVM, the CernVM File System and pilot job environments such as DIRAC or PANDA, offers the exciting possibility of delivering a clean separation between the roles of sites—to deliver computing resources to competing virtual organisations according to local policies—and virtual organisations—who need to allocate resources provided across a computing grid amongst their members according to the VO’s changing priorities. Exploiting these different developments effectively would allow the High Energy Physics community (or any other) to combine the best features of the Grid and Cloud computing models—or, in other words, to build a Grid of Clouds.

[1] S Andreozzi et al; 2008 J. Phys.: Conf. Ser. 119 062011, “Towards the integration of StoRM on Amazon Simple Storage Service (S3)”
[2] Martin Sevior et al; 2010 J. Phys.: Conf. Ser. 219 012003, “Belle monte-carlo production on the Amazon EC2 cloud”
[3] e.g. the JSPG Traceability and Logging Policy at http://www.jspg.org/wiki/JSPG_Docs
[4] https://www.hepix.org/
[5] http://www.hpc2n.umu.se/events/workshops/09/hepix/ and http://indico.cern.ch/conferenceDisplay.py?confld=45282
[6] http://indico.cern.ch/conferenceDisplay.py?confld=61917

Figure 1: Illustration of how trusted machine images could be integrated with pilot job frameworks to provide a cloud-like computing environment across sites in a computing Grid. (Illustration courtesy of Ulrich Schwickerath of CERN.)
[7] w3.hepix.org/virtualization
[8] http://www.jspg.org/wiki/Policy_Trusted_Virtual_Machines
[9] https://bitbucket.org/rpathsync/amiconfig
[10] hepixvmilssubscriber at https://github.com/hepix-virtualisation/hepixvmilssubscriber
[11] http://stratuslab.eu/
[12] http://glance.openstack.org/
[13] P Buncic et al.; The European Physical Journal Plus 126(1) 1–8 10.1140/epjp/i2011-11013-1, "A practical approach to virtualization in HEP"
[14] J Blomer et al.; 2011 J. Phys.: Conf. Ser. 331 042003, "Distributing LHC application software and conditions databases using the CernVM file system"