Running Agilent GeneSpring MPP on the Cloud

Technical Overview

Authors

Stephen Madden, Rick A. Fasani,
and Michael Rosenberg

Agilent Technologies, Inc.

Santa Clara, California, USA

Introduction

Cloud computing means efficiently sharing a pool of interconnected computational resources such as processing power, disk space, network bandwidth, and software applications among users as they need them. Rather than outfitting each user with an individual desktop PC including these elements, the user can connect to a remote virtual machine (VM) to perform their work. The centralization and efficient use of resources offers economies of scale, and third-party providers of cloud services allow customer organizations to take advantage of a pay-as-you-go structure to decrease the costs of computing. In addition, cloud computing strongly facilitates collaboration. Users can easily share data sets loaded in a shared application running on a central virtual server.

The Agilent GeneSpring MPP Platform is designed to be easily hosted in a cloud computing environment. Beginning with version 13, support has been added to run effectively on a VM, a key technology in the cloud.

Agilent Technologies

Types of Virtualization

GeneSpring MPP supports three types of machine virtualization:

- **Personal:** VMs on a local laptop or workstation

- **Private:** VMs on internal servers running behind the user's firewall (internal data center)

- **Public:** VMs running on secure remote servers located outside the user's firewall (cloud)

Personal refers to VMs a user creates and manages themselves on their own machine. *Private* usually refers to a group of shared VMs centrally managed by an IT department, and often called a private cloud. In both cases, the physical hardware is owned by the consuming organization. *Public* generally refers to when another organization owns and administers the hardware, and leases out secure access to its customers.

A **personal VM** is usually created on a computer assigned to a user with VM software installed such as VMWare, Virtual PC, or VirtualBox. The VM software runs on top of the host operating system as a software program. It allows the user to create, clone, backup, start up, and shut down multiple VMs running either a copy of the native operating system (OS) (for example, Linux) or another OS (such as Microsoft Windows 7) (Figure 1). This gives the user the convenience of running application software in a specialized environment, for example, if the software program would only run on a Windows operating system. This flexibility is made possible by full hardware virtualization, in which the software emulates the underlying hardware of a VM.

A **private cloud** is hosted on a relatively powerful computing infrastructure set up by the organization for multiple users. The hardware could be located in the same physical location as the users, or accessible remotely through the organization's intranet. Each user may be provided their own remote VM to run application software such as GeneSpring MPP. To manage the VMs on the internal cloud, IT staff uses enterprise-level virtualization software such as Xen, VMWare ESX, or Microsoft Hyper-V.

Remote computing services are usually referred to as *cloud computing*, or being *on the cloud*. Examples are Amazon Web Services (AWS), Google Cloud Platform, and Microsoft Azure. These vendors host large server farms and are able to provision virtual machines for their customers on-demand, with the desired processing power, memory, and file storage space. Customers only pay for what they use, which gives them flexibility and, in many cases, cost savings.

	Windows application software (for example, MS Office)
	Windows 7
Linux application software (for example, Terminator for terminal emulation)	Virtual Machine (VM)
Linux	
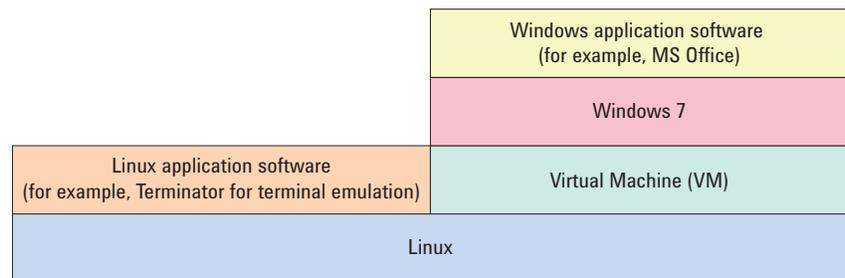
Figure 1. An example architecture with VM software allowing the user to run application software from another operating system.

Configuring GeneSpring MPP for Cloud or VM Deployment

The steps to configure GeneSpring MPP for use with VMs are:

1. Install the GeneSpring MPP software on a VM according to standard installation instructions.

2. (Optional) Clone the VM for the desired number of individual users.

3. Start up the VM using the instructions provided with the VM software, by your administrator, or by your remote provider.

4. Connect to the VM using a program such as Microsoft Remote Desktop Connection.

5. Start the GeneSpring MPP software. The License Manager dialog will appear during the first start up (Figure 2). Enter the Order IDs for the products purchased, select the **Cloud or VM Installation** checkbox, then click **OK**.

6. Repeat Steps 3 through 5 for each cloned VM with a new set of Order IDs.
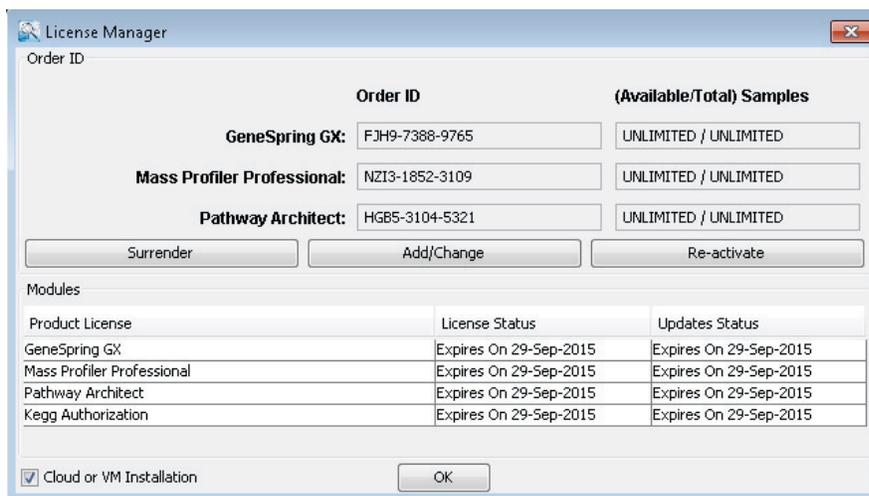
Figure 2. The Agilent GeneSpring MPP License Manager dialog (available from the Help → License Manager menu item) with the **Cloud or VM Installation** checkbox selected.

Note: for each individual VM, the desired Agilent GeneSpring, MPP, or Pathway Architect products need to be purchased and Order IDs obtained. Order IDs cannot be reused on different VMs. Also, to authenticate the Order IDs, the VM needs access to the Internet whenever GeneSpring MPP is activated (Step 5), and whenever it is started during normal operation.

Multiple VMs can be configured to access the customer's existing file-sharing infrastructure (shared drives).

This is one of the most compelling aspects of cloud computing in scientific contexts—as data sets increase in size, it becomes more difficult to download data to a local machine. Centrally hosted VMs "take the software to the data", giving users a central place for performing operations. There is no need to copy massive file sets to their local machine to perform operations locally. GeneSpring MPP makes it easy to share projects between users as tar files as well as with file-sharing services such as Dropbox and Google Drive.
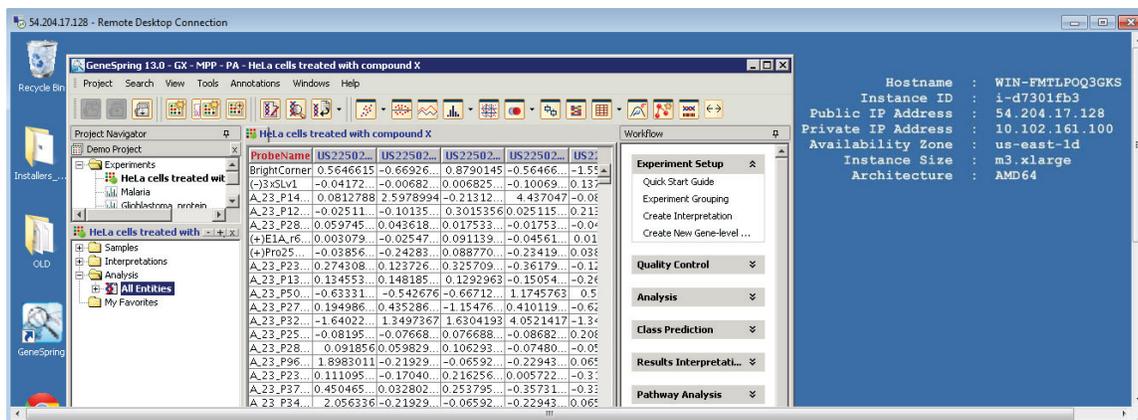

Figure 3. Agilent GeneSpring MPP running on Amazon Web Services (AWS). Note the server specifications in the upper right corner. Multiple servers can be provisioned on-demand, demonstrating the flexibility that cloud computing provides.

Considerations for Cloud or VM Installations

Although virtualization and cloud computing can streamline or eliminate many common IT issues, multiple VMs with GeneSpring MPP installed still present some of the traditional IT challenges associated with managing desktop deployments:

- **User management**
  A knowledgeable person is required to configure user accounts and permissions as needed for all VMs.

- **Software license management**
  Each instance of GeneSpring MPP, whether installed directly on a VM, or copied on a duplicate VM, requires a separate set of Order IDs obtained from Agilent and then entered into the VM instance's License Manager dialog.

- **Software version management**
  As Agilent releases new versions of GeneSpring MPP, a decision must be made whether to obtain and install the new version on each VM.

- **Backup of GeneSpring MPP content**
  The backup process has to be enabled on each VM. This can be done before or after the VMs are cloned. Storing multiple copies of backup in a public cloud may be expensive; to optimize storage costs, users should select the least expensive long-term storage option, as described in the next section.

Virtualization and cloud computing also present new challenges and considerations:

- **Cloud administration**
  Knowledge of the specific cloud infrastructure selected by the customer (for example, AWS) is necessary to configure, provision, and maintain cloud VMs.

- **Optimizing computing costs**
  While cloud computing can provide flexibility with its pay-as-you-go model, most remote computing services charge while the VM instance is running. If no user is logged on, the organization accumulates charges for unused computing time. Therefore, it is important to either educate users about stopping unused VMs, or deploy software tools that monitor resource usage and stop unused servers automatically.

- **Optimizing storage costs**
  Remote computing services also charge for disk storage space. Organizations need to manage this as users continue to upload data. In some cases, users might unknowingly upload data files and projects that are already available to them in the cloud. Careful management of storage space, including the use of file sharing and non-duplication, can create significant cost savings for organizations.

Case Study: The Human Toxome Collaboratorium

The Human Toxome Collaboratorium is an example of a successful deployment of a cloud computing environment for academic research. As part of a long-term vision to modernize toxicity testing for the 21st century[1,2], a consortium of six academic and commercial organizations partnered to map the pathways of toxicity using endocrine disruption as a model system[3,4]. Data were generated and analyzed at multiple sites using various technologies including Agilent gene expression microarrays, Next-Generation Sequencing (NGS), and Agilent Liquid Chromatography/Mass Spectrometry (LC/MS). The Human Toxome Collaboratorium is a set of shared VMs, hosted on AWS, that contain a full complement of software

applications already installed, licensed, and configured. The VMs are backed by virtual disks, or hard drives, that contain relevant public and private data, as well as the free space necessary for new data and analyses. Figure 4 depicts the overall architecture, grouped by -omics technology. The primary interface to the Collaboratorium is a shared desktop, for which each user has an account that can be accessed through the Remote Desktop Protocol (RDP) from any Windows, OS X, or Linux machine.

To reduce costs, relevant VMs have been configured to shut down automatically after an extended period of inactivity. Users can restart the VM when they need it. In addition, the consortium has eliminated free space and large duplicate data sets, which has reduced the total data storage costs by approximately 50 %.

Conclusions

Running Agilent GeneSpring MPP on a virtual machine, either on the desktop or in the cloud, can offer many advantages in terms of collaboration, flexibility, and cost savings. The software is easy to install and configure on multiple VMs, which can then be provided to multiple users. However, organizations must be aware that traditional IT challenges do not disappear in the cloud, and may subtly change. Nevertheless, when successfully deployed, a cloud configuration can increase research productivity and decrease costs, thus accelerating the pace of scientific discovery.
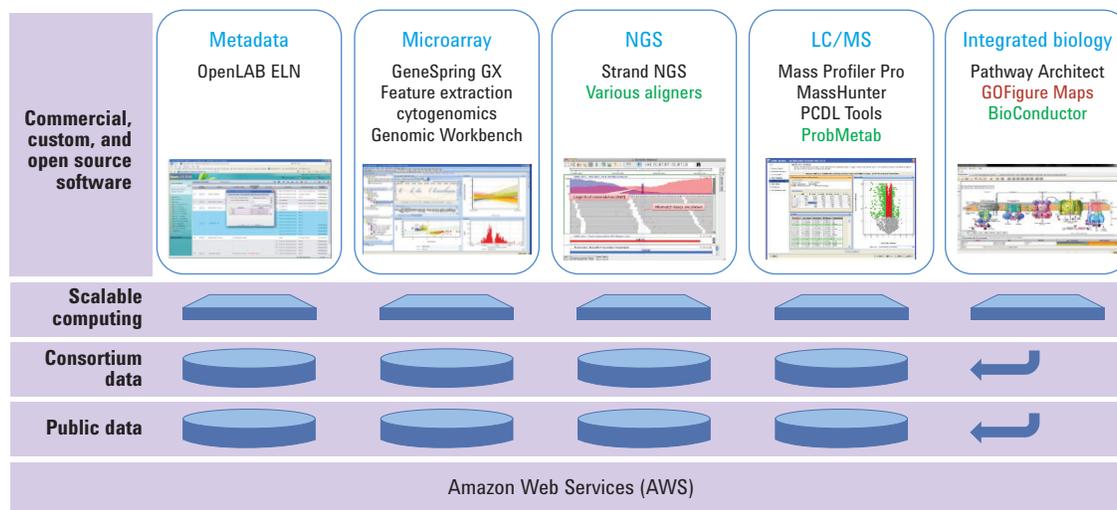

Figure 4. Architectural overview of the Collaboratorium. Horizontal layers indicate fundamental components, including commercial, open source, and custom-built software (blue, green, and red lettering respectively). Vertical columns show the software and data grouped by various -omics technologies.

5

Acknowledgements

References

1. National Research Council. *Toxicity testing in the 21st century: a vision and a strategy*; National Academy Press; Washington, DC, 2007.

2. Stephens, M. L; *et al.* Evidence-based toxicology for the 21st century: opportunities and challenges. *ALTEX* **2013**, *30(1)*, pp 74-103.

3. Hartung, T; *et al.* Food for Thought... Systems Toxicology. *ALTEX* **2012**, *29* pp 119-128.

4. Hartung, T; *et al.* Food for Thought... Integrated Testing Strategies for Safety Assessments. *ALTEX* **2013**, *30* pp 3-18.

For more information

For more information on our products and services, visit our Web site at www.agilent.com/chem.

www.agilent.com/chem

Agilent Technologies