NEP-101 System Design Specification Document

NEP-101 Team

March 11, 2014

Version 0.1

Revision History

Name	Date	Reason for Changes	Version
Colin Leavett-Brown	10/28/2013	Skeleton	0.1

Contents

1 Introduction

The use of cloud computing technologies is becoming ubiquitous for commercial and research computing. The HEP Legacy Data project (NEP-52) contributed to this trend, establishing a distributed cloud computing environment particularly suitable for compute intensive workloads. For the last several years, researchers within both particle physics and astronomy have used this distributed cloud model to great effect. However, they have been limited to workloads with modest input requirements. The NEP-101 HEP Data-Intensive Distributed Cloud Computing project will expand the capability of the distributed cloud by enabling it to run data intensive applications. Our goal is to run ATLAS jobs that can process real or simulated data samples that require up to 20 GB per core in a 12-hour period. Although the storage per core is modest, the amount of data required could reach 40 TB per day if the system is running 1000 or more simultaneous jobs. Enabling the system for data-intensive applications requires the addition of new functionality in the existing services and the integration of new technologies or services. We plan to use a federated system for managing distributed data. ATLAS already has its data distributed over 100 centres around the world. Our goal is to have each job either stream or stage the data from the nearest centre.

1.1 Purpose

The purpose of this document is to present the system architecture, components, design details and testing environments for the **NEP-101 HEP Data-Intensive Distributed Cloud Computing** project. It is intended as a guide for current and future developers, and will be delivered to CANARIE as a work product of the NEP-101 project.

1.2 Scope

The key component of the distributed cloud computing system is the Cloud Scheduler service, which was developed in an earlier CANARIE NEP project. Cloud Scheduler manages application jobs and VM images over the distributed cloud. We will need to modify the Cloud Scheduler service so that it can become data-aware while making its scheduling decision.

In addition, new technologies have been developed that would significantly enhance the operation of the distributed cloud system. For example, in 2012 OpenStack has become the most popular cloud infrastructure software and approximately half of the clouds used in our system use OpenStack. In this project we will modify Cloud Scheduler so that it can use the OpenStack application interface and enhance the functionality of our VM image repository so that it can automatically distribute the images to all cloud types.

The distributed cloud uses a single Squid cache for accessing the application software. A better and more reliable method would use a central server that points the cloud to the nearest Squid cache. We have a prototype system under evaluation and plan to use it in the production system.

Further, we are very interested in exploring the use of micro-VM images. Micro-VMs have the potential to reduce the size of the VM images by three orders of magnitude to tens of megabytes. Micro-VMs have the operating system software in the same remote

software caches as the application software. This would radically impact the way we view and manage VM images.

1.3 Definitions, acronyms and abbreviations

APF	The ATLAS AutoPyFactory generates batch jobs in response to workloads within the ATLAS PanDA queue. These APF batch jobs are payloadless, that is they do not contain the application or data to perform ATLAS analysis or simulation, but they are used to secure computing resources in a distributed grid or cloud environment. Having secured the required computing resources, an APF job selects and pulls a unit of work from the PanDA queue for execution.
ATLAS	A HEP experiment at CERN laboratory in Geneva, Switzerland (http://atlas.ch)
Cloud	A collection of computing resources and software that provide on demand through Web Services, Infrastructure As A Service (IaaS) Virtual Machines (see below).
DCCM	Distributed Cloud Computing Model as developed by NEP-52 and enhanced by NEP-101.
HEP	High Energy Physics, sometimes referred to as Particle Physics or Nuclear Physics
Image Repository	A storage capability employed by the NEP-101 system to manage virtual machine images. Depending on its' type, an image repository may be dedicated to a single cloud or shared by multiple clouds.
NEP-52 [1]	HEP Legacy Data project (Oct 2009 - Mar 2012), created the DCCM for high throughput, modest data input/output serial processing.
NEP-101	HEP Data-Intensive Distributed Cloud Computing project (Oct 2013 - Dec 2014), will extend the DCCM for data intensive ATLAS applications.
Virtual Machine (VM)	A complete computer system represented in software

1.4 References

1.5 User Perspective

The system developed will extend the services provided by the distributed cloud model to facilitate the processing of ATLAS production jobs requiring an order of magnitude more input data than is currently practical. This will more than double the class of jobs suitable for this computing environment. These services include both interactive and batch processing capabilities providing enhanced facilities for the reading and writing of large data sets, the management of network traffic through advanced caching techniques, and the optimization, management and distribution of VM images.

[1] https://wiki.heprc.uvic.ca/twiki/bin/view/Main/CanarieProjectNEP52

Batch services, data federation, closest repository, WebDAV/FAX.

Batch services, CVMFS, Shoal/Squid.

Interactive services are extended to manage the distribution of VM images. Previously, researchers were empowered to create and upload images to the repository, to list and instantiate (launch) images within the repository that they either own (created by them) or were shared by another user. Having launched an image the user may create a new image by logging into into the running image as the super user, modifying the image to their requirements, and saving a personal copy or snapshot of the modified image. Through the use of a web browser, the NEP-101 extensions will allow the researcher to distribute available images to any cloud for which they have credentials, either manually or automatically through the creation of distribution profiles. Additionally, the owner of an image may rename or delete a distributed image, or revoke the privileges of other users to use it.

MicroKernel CernVM.

In every case, users must authenticate with each component of the services provided.

1.6 Document Organization

The rest of this document is organized as follows. Section **??** gives an overview of the functionality of the system. It describes the general structure of the system and its informal requirements. It is intended for a general audience. The remainder of this document, sections **??**, **??**, and **??**, describes in technical terms the details of the functionality of the system as well as the constraints imposed on it. These sections are intended for developers.