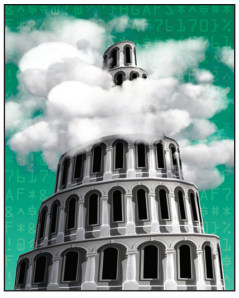
Editor: George Pallis • *gpallis@cs.ucy.ac.cy*

Experimenting with Next-Generation Cloud Architectures Using CloudLab



Aditya Akella • *University of Wisconsin–Madison*

The infrastructure presented here (called CloudLab) is distributed, open, and programmable to enable deep experiments on new cloud architectures and the applications that they will enable.

The computer science community is bubbling over with a wide variety of creative ideas for next-generation cloud architectures, ranging from improvements to today's OpenStack software suite to radical new architectures embracing new storage paradigms, green clouds, distributed clouds and cloudlets, clouds geared toward efficient support for mobile devices and the Internet of Things, privacy-preserving clouds, and so on. Unfortunately, today's cloud computing environments don't enable research into the very architectural elements that enable clouds themselves. Many of the ideas that drive modern cloud computing, such as server virtualization, network slicing, and robust distributed storage, arose from the research community. But because today's clouds have these ideas "baked in," they're unsuitable as facilities in which to conduct research on future cloud architectures or to explore radical new ideas that run counter to today's prevalent wisdom.

To enable foundational cloud research, we need infrastructure that can support research into a wide variety of cloud architectures, ranging from variations of today's architectures to radically new "clean slate" cloud concepts. This infrastructure must be well-instrumented, transparent, and capable of supporting new architectures fundamentally based on energy efficiency or mobile and cyber-physical applications, and into Big Data and new storage paradigms. Keeping this in mind, here I present CloudLab Berkeley Software Distribution (see http://cloudlab.us), a facility for transformative research into new cloud architectures and the applications that they will enable.

CloudLab

CloudLab isn't a cloud: rather, it provides the substrate on which researchers can build their own clouds and experiment in an environment that provides a high degree of scientific fidelity. It's flexible enough to enable research into evolutionary and revolutionary changes to cloud architecture. We hope that these advances in cloud architecture will facilitate new classes of applications not possible today. CloudLab has an open access use policy, and will be available without charge to researchers and educators nationwide.

CloudLab provides architectural flexibility on two fronts: the hardware that comprises the facility, and the software (including cloud stacks and applications) that will be capable of running on it. CloudLab's approach to hardware flexibility focuses on diversity and configurability.

It consists of three large-scale datacenters located at Clemson (built in collaboration with Dell), Utah (HP), and Wisconsin (Cisco), with each making different architectural choices with regard to hosts, networks, and storage. Dimensions of diversity include software-defined networking capabilities (SDN; see http://opennetworking.org), storage technologies, network topologies, and processor architectures. CloudLab datacenters are interlinked via 100-gigabit per second (Gbps) IP and Layer 2 connections on Internet2's Innovation Platform, and experiments may incorporate multiple datacenters as desired.

To enable software flexibility, CloudLab's central feature is a control framework that operates at a lower layer than cloud software stacks: it directly provisions and controls "raw" hardware.

 1089-7801/15/$31.00 © 2015 IEEE Published by the IEEE Computer Society

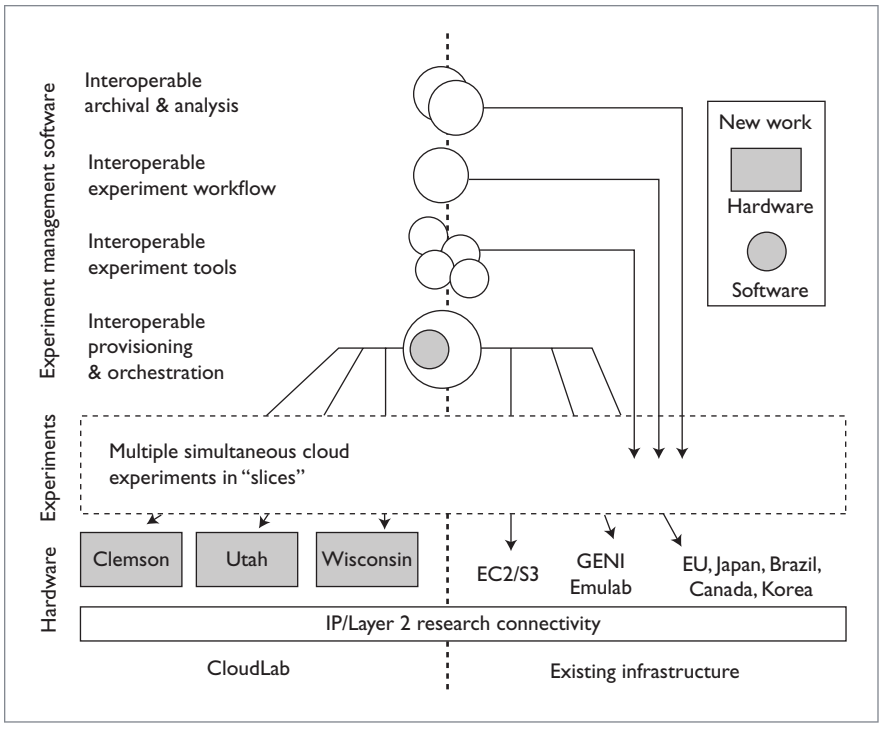


Figure 1. CloudLab's components. CloudLab uses IP and Layer 2 services from national and regional research networks (such as Internet2) that are directly compatible with the Global Energy Network Institute (GENI).

Operating the infrastructure at this low level allows researchers tremendous freedom to experiment with the higher layers: while CloudLab will provide "canned" configurations for quick installation of popular cloud stacks, storage systems, and computational frameworks, researchers aren't bound to any of these, and are free to build whatever they wish on top of the physical resources provided by CloudLab. Researchers who build clouds in CloudLab will eventually have the option of evaluating their work with synthetic workloads that we provide, or opening them up to use by actual applications by making their architecture available as an image within that application. CloudLab's design is such that reprovisioning occurs on the order of minutes, allowing it to flexibly support experiments in the range of hours to years.

A key factor distinguishing CloudLab from commercial cloud offerings such as Amazon's Elastic Compute Cloud (EC2) and Azure is that it's designed first and foremost as a scientific instrument. CloudLab provides a strong degree of isolation between simultaneous experiments, ensuring that results gathered on it won't be tainted by concurrent experiments. It's transparent, providing users with the details about the necessary details of infrastructure so that they can rigorously evaluate the systems that run on top. CloudLab provides deeper controllability — experimenters can control the entire software stack on individual compute nodes, they can program the network interconnect, and they can leverage all levels of storage available per node — enabling innovative and meticulous studies, such as low-latency datacenter networking (akin to Mohammad Alizadeh and his colleagues' work[1]), RAMClouds,[2] and nested virtualization.[3] Currently, we're working on heavily instrumenting it, so as to enable cloud architects and application developers to not only observe high-level effects, but to analyze and understand their low-level causes. In addition, we're working on mechanisms for letting experiments introduce controllable degraded operation, emulating failures, performance anomalies, and fluctuations in power availability.

Architecture

Figure 1 depicts the major components of CloudLab. Reading from the bottom of Figure 1 and going upward, CloudLab uses IP and Layer 2 services from national and regional research networks (such as Internet2) that are directly compatible with the Global Energy Network Institute (GENI). At the hardware layer, CloudLab has a three-site distributed cloud infrastructure in Phase 1, with datacenters at Clemson, Utah, and Wisconsin.

Building on Emulab and GENI, CloudLab supports large numbers of simultaneous experiments in its sliced infrastructure, as Figure 2 shows. Experiments may be contained within a single site, or extend across multiple sites. We can support at least 100 simultaneous experiments at each site (at 50 CPU cores each), or single large experiments of up to 15,000 cores and 1 petabyte of storage within CloudLab proper. Augmenting these experiments with commercial offerings such as EC2 allows for extremely large-scale experiments, although EC2 segments don't provide CloudLab's transparency and thorough instrumentation.

CloudLab Hardware Details

As we mentioned, at its core, CloudLab is a distributed, three-site infrastructure. Each site comprises approximately 5,000 cores and 300–500 terabytes of storage in the latest virtualization-capable hardware. We provide 2 × 10-Gbps networking to every node via SDN (such as OpenFlow; see http://archive.openflow.org). A 100-Gbps full-mesh SDN interconnect lets researchers instantiate a wide range of in-cluster experimental topologies — for example,

fat trees, rings, hypercubes, and so on. We provide two major types of storage: per-server storage (a mix of high-performance flash and high-capacity magnetic disks at a ratio of about one disk per every four cores), and a centralized storage system. This storage mix enables a range of experiments with file systems, storage technologies, and Big Data, while providing convenient, reliable file systems to researchers who aren't interested in storage experiments.

The University of Wisconsin–Madison is partnered with Cisco Systems to build a powerful and diverse cluster that closely reflects the technology and architecture used in modern commercial datacenters. Figure 3 illustrates the cluster. The initial cluster has 100 servers with a total of 1,600 cores connected in a Clos fat-tree topology. Future acquisitions in 2015 and 2016 will grow the system to at least 240 servers.

Currently, the servers are broken into two categories, each offering different capabilities and enabling different types of cloud experiments. In the initial cluster, all servers will have the same CPU (2 × 8 cores at 2.4-GHz), RAM (128 Gbytes), and network (2 × 10 Gbps to top-of-rack, or ToR) configuration, but will differ in their storage configurations. Each of the 90 servers in the first category will have 2 × 1.2-Tbyte disks. We expect these to be used for experimenting with exciting new cloud architectures and paradigms, management frameworks, and applications. Each of the 10 servers in the second category will have a larger number (1 × 1 Tbyte, 12 × 3 Tbytes donated by Seagate) of slower disks. This category is targeted toward supporting experiments that stress storage throughput. Each server will also have a solid-state drive (SSD; 480 Gbytes) to enable sophisticated experiments that explore storage hierarchies in the cloud.

The servers use Nexus switches from Cisco for ToR switching. Each ToR Nexus is connected to six spine switches via dedicated 40-Gbps links.

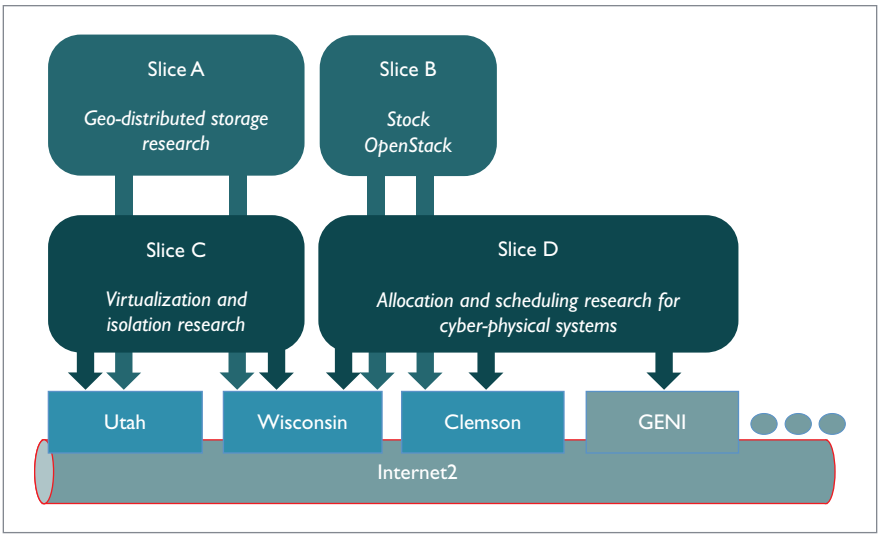


Figure 2. CloudLab has a sliced infrastructure that supports large numbers of simultaneous experiments. It flexibly allows experiments to be contained within a single site, or to extend across multiple sites.

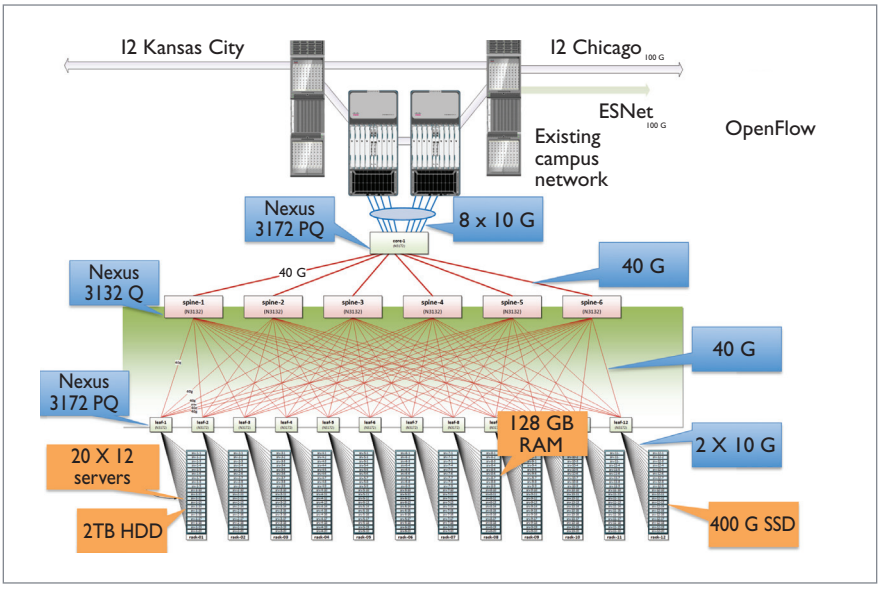


Figure 3. CloudLab's cluster, built to closely reflect the technology and architecture used in modern commercial datacenters. The initial cluster has 100 servers with a total of 1,600 cores.

Each spine will connect via a 40-Gbps link to a Nexus wireless area network (WAN) switch for campus connectivity to Internet2 and the other two CloudLab facilities. We selected the Cisco Nexus series because it offers several unique features that enable broad and deep instrumentation, as well as a wide variety of cloud networking experiments. Examples of these features include OpenFlow 1.0; monitoring instantaneous queue lengths in individual ports; tracing control plane actions at fine time scales; and support for a wide range of routing protocols.

The remaining two clusters at Utah and Clemson are built according to a similar reference architecture as the Wisconsin cluster, but they differ in the hardware details, and hence, in the nature of the experiments that they enable.

The University of Utah is partnering with HP to build a cluster with 64-bit ARM processors and OpenFlow 1.3 support throughout. This cluster will consist of 7 HP Moonshot Chassis, each having 45 eight-core ARM servers (315 servers, 2,520 cores total) with 64 Gbytes of RAM (20 Tbytes total), and 120 Tbytes of serial AT attachment (SATA) flash storage (38 Tbytes total). Each chassis has two ToR switches, and each server has two 10-Gbyte network interface controllers (NICs), one connected to each of the ToRs. Each ToR has 4 × 40 Gbytes of uplink capacity to a large core switch, for a total of 900 Gbits of connectivity within the chassis and 320 Gbits of connectivity to the core. One option for allocation will be to allocate an entire chassis at a time; when allocated this way, the user has complete administrative access to the ToR switches in addition to the nodes. Users allocating the entire chassis will also be given administrative access to a "slice" of the core switch using Multitenant Device Context (MDC), which gives the user a complete virtual switch, including full control over Layer 2 and 3 features and a dedicated OpenFlow datapath.

The Clemson system, developed in close cooperation with Dell, has three major components: bulk block storage, low-density storage for MapReduce/Hadoop-like computing, and generic VM nodes used to provision virtual machines. All nodes have 16 cores per node (2 CPUs), one on-board 1-Gbit Ethernet and a dual-port 56/40/10-Gbps card. Bulk storage nodes provide block-level services to all nodes over a dedicated 10-Gbps Ethernet. Storage nodes consist of 12 × 4-Tbyte disk drives in each node, plus 8 × 1-Tbyte disks in each node. The nodes are configured with 256 Gbytes of memory. Hadoop nodes have a 4 × 1-Tbyte disk and 256 Gbytes of memory. The VM nodes each have 256 Gbytes of memory. The large memory configuration reflects the need for significant memory in VMs today and lets us increase performance by reduced paging/swapping in the VMs.

CloudLab Software Details

The software stack that manages CloudLab is based on Emulab, a testbed control suite developed at the University of Utah. Exactly as they do in Emulab today, researchers or students start a CloudLab experiment by specifying what resources they need via an easy-to-use graphical Web interface and scriptable commands. This specification works at two levels: a description of the physical resources for the experiment (the set of compute, network, and storage resources) and the software that should be run on top of those resources (disk images for compute nodes and controllers for SDN infrastructure). CloudLab provides a menu of "canned" specifications for users to get started quickly, and experimenters will also be able to create their own by modifying the canned specifications or starting from scratch. For example, one canned cloud specification might run OpenStack "Havana" (see www.openstack.org/software/havana) to use KVM for virtualization, with a recent Linux OS image, run on a fat-tree network topology, and use the Floodlight controller on OpenFlow. Another might use CloudStack 4.2, a simple-tree topology, Internet Small Computer System Interface (iSCSI)-based storage, the Xen hypervisor with FreeBSD, and plain Ethernet switching.

CloudLab users will be in no way limited to these standard software stacks: with bare access to the hardware, any operating system, hypervisor, storage stack, SDN control plane, or cloud-control system supportable on CloudLab's hardware is possible.

To achieve the fidelity of a scientific instrument, each cloud instantiated in CloudLab uses a distinct set of hosts, and in most cases, network devices and storage. To the extent that network or storage may be shared, CloudLab makes this visible, and monitors shared resources so that experimenters can understand the effects sharing might have on their results. By providing isolation at the base layer, CloudLab frees users to experiment with different types of isolation at higher layers — for example, some clouds in CloudLab may provide best-effort service, while others may experiment with real-time guarantees.

CloudLab uses two primary mechanisms for resource provisioning: imaging for compute and storage resources and SDN control for network resources. Imaging is accomplished with the Frisbee disk-loading system created for Emulab, which is extremely fast and highly scalable. SDN control (see http://opennetworking.org) is accomplished by allowing experiments to run their own SDN controllers. A fallback to regular Ethernet forwarding will also be available for experiments that don't require SDN.

Once a cloud is running in CloudLab, it has the option of using a set of infrastructure services that CloudLab will provide, or ignoring those services completely. Infrastructure services include account management, direct access to the imaging and network-provisioning systems, storage provisioning, SDN control, and helper support for running higher-level stacks such as cloud-control stacks and HPC schedulers.

CloudLab is capable of introducing precise, controllable, emulated fault injection, another feature it inherits from Emulab. The faults can include network degradation (link-down events, bandwidth limitations, increased latency, and random packet loss), host failure, and disk problems (read/write errors and increased I/O latency). This degradation will be

implemented by interposing on the network and storage stacks, although this interposition only occurs if explicitly requested. It's controllable at runtime, giving different, controllable characteristics over an experiment's life.

With the increased use of cloud computing, it's time to ask how we enable research in building future clouds that help us develop hitherto unforeseen applications and use cases. Existing commercial infrastructures don't provide the low-level control and instrumentation necessary to conduct this research. With its distributed, open, programmable infrastructure, CloudLab bridges the gap to allow rich, deep experiments on cloud architectures. Parts of CloudLab are still being pieced together, but the core infrastructure exists and is already being used by many tens of research groups and experimenters.

References

1. M. Alizadeh et al., "Less Is More: Trading a Little Bandwidth for Ultra-Low Latency in the Data Center," *Proc. 9th Usenix Conf. Networked Systems Design and Implementation*, 2012, p. 19; www.usenix.org/conference/nsdi12/technical-sessions/presentation/alizadeh.
2. J. Ousterhout et al., "The Case for RAMClouds: Scalable High-Performance Storage Entirely in DRAM," *SIGOPS Operating Systems Rev.*, vol. 43, no. 4, 2010, pp. 92–105.
3. M. Ben-Yehuda et al., "The Turtles Project: Design and Implementation of Nested Virtualization," *Proc. 9th Usenix Conf. Operating Systems Design and Implementation*, 2010, pp. 1–6.

Aditya Akella is an associate professor in the Computer Science Department at the University of Wisconsin–Madison. His research interests focus on computer networking, spanning the areas of network architecture, Internet routing, network management, measurement, and network security and wireless networking. Akella has a PhD in computer science from Carnegie Mellon University. Contact him at akella@cs.wisc.edu.