Koala: A Discrete-Event Simulation Model of Infrastructure Clouds

Koala is a discrete-event simulator that can model infrastructure as a service (IaaS) clouds of up to $O(10^5)$ nodes. The model, written in SLX[1], facilitates investigation of global behavior throughout a single IaaS cloud. Koala is uniquely positioned between microscopic models, which require hundreds of parameters to fully specify detailed behavior of individual cloud elements, and macroscopic models, which aggregate behavior of many cloud elements into abstract representations that may lose key interactions on fine spatiotemporal scales.

Koala is based loosely on the Amazon Elastic Compute Cloud (EC2) and on Eucalyptus open-source cloud software. ***Koala*** simulates four commands (RunInstances, DescribeInstances, RebootInstances and TerminateInstances) from the EC2 application programming interface (API) and also simulates seven Virtual Machine (VM) types inspired by EC2 offerings. ***Koala*** simulates cloud controller, cluster controller and node controller components modeled after Eucalyptus, and all controllers communicate using Web Services, as is the case in Eucalyptus. ***Koala*** modifies the design of Eucalyptus in the following ways: (1) supports multiple VM types in a single request (as inferred from the EC2 API), (2) avoids centralization of node information at the Cloud Controller (in order to support scalability) and (3) allows resource allocation to proceed partially in parallel (reducing queuing delays during periods of intense demand).

Koala models infrastructure clouds in five layers: (1) demand layer (including user behavior), (2) resource allocation layer (including alternate resource allocation algorithms), (3) supply layer (including specification of configurations for node platforms), (4) Internet/Intranet layer (including the ability to locate cloud clusters remotely from the cloud controller) and (5) an optional VM behavior layer (including the ability to simulate VM behavior with or without simulating tasks). Resource allocation algorithms, implemented in the cloud controller (for selecting clusters) and in each cluster controller (for selecting nodes), include random, least-full first, first-fit, next-fit, most-full first, percent allocated and tag-and-pack.

Koala can record hundreds of measures reflecting behavior and dynamics at the user, the cloud, the cluster, the node, the Internet/Intranet and the VM behavior layers. Further, ***Koala*** can record temporally varying behavior for each node and user in the simulation, providing the ability to animate and visualize spatiotemporal behavior in a cloud simulation.

Initial sensitivity analyses of the simulator reveal that ***Koala*** is driven primarily by three parameters: (1) number of users, (2) number of clusters and (3) number of nodes per cluster. The number and types of VMs requested by users and the configuration of node platforms within a cloud also exhibit modest influence on ***Koala*** behavior. Minor influence on ***Koala*** behavior was seen for the duration that users hold their VMs and the algorithm (least-full first or percent allocated) used by the cloud to allocate VMs to clusters. ***Koala*** was also found to exhibit five main behavioral characteristics: (1) congestion, (2) cloud-wide resource usage, (3) variance in cluster load, (4) mix of VM types and (5) number of VMs running in the cloud.

---

[1] SLX is a commercial simulation system available from Wolverine Software, See http://www.wolverinesoftware.com