GSAW 2010
Working Group Session 11D

Eucalyptus-Based Event Correlation

Nehal Desai
Member of the Tech. Staff, CSD/CSTS/CSRD, The Aerospace Corporation

Dr. Craig A. Lee, lee@aero.org
Senior Scientist, CSD/CSTS/CSRD, The Aerospace Corporation
President, Open Grid Forum

March 3, 2010
Overview

• Goal:
  – Prototype and evaluate a Cloud Computing Environment as a generic hosting environment for NSS applications

• Approach:
  – Augment an existing analyst tool, CORE, with a Correlation Evaluator (CE) that is dynamically provisioned and run in a prototype private cloud
    • CE will automate and enhance the semantics and scope of correlation queries against the a database aggregator to identify causal events of the highest importance
    – Use prototype environment to quantitatively evaluate benefits of cloud computing

• Success Metrics:
  – Demonstrate improved server utilization in the private cloud
  – Demonstrate ability to dynamically scale-out support for multiple CEs and CORE users
  – Demonstrate improved ability to identify high impact causal events
  – Demonstrate ability to be a generic hosting environment
Approach Strategy

• Phased development
  – *Build critical functionality first*
    • Demonstrate simplest possible test cases as early as possible
  – *Add functionality incrementally to ultimately demonstrate complete system*
    • Ability to dynamically support multiple users, across multiple sites, hosting multiple applications
    • Automatically enforce data policy across sites

• Assess project results at every increment
  – *Be a "fast learning" project that can quickly follow successful efforts and discard disappointing ones*

• Leverage existing hardware platforms
Major Prototype Components

- **CORE**
  - Google Earth-based User interface with "tree" of available data

- **Database Aggregator (DA)**
  - Provides access to multiple databases

- **Eucalyptus**
  - Open source Amazon EC2 API clone to be used for private cloud

- **Correlation Evaluator Client (CEC) (new development)**
  - Panel added to CORE user interface to specify correlation queries

- **Correlation Evaluator (CE) (new development)**
  - Cloud application that runs correlation queries from one or more CECs

- **iRODS**
  - Service for managing distributed data archives w/ integrated Rule Engine

- **Workflow Manager (WfM)**
  - Tool to manage sets of queries against the DA

- **Performance Model**
  - A model to estimate the "cost" of doing queries before they are done
• **Elastic Utility Computing Architecture Linking Your Programs To Useful Systems**
  - Open Source API Clone of Amazon EC2
  - Web services based implementation of elastic/utility/cloud computing infrastructure
  - University of California, Santa Barbara
• $5.5M in venture capital secured
  - Intends to be the “Redhat” of cloud computing
Illustration of CORE with Correlation Evaluation Client
Correlation Evaluator (CE)

- Correlation Evaluators (CEs) take spatial/temporal queries from CECs
  - Checks for partial/component correlations already managed by iRODS (see below)
  - Run sets of workflows against the DA to derive correlations
    - Dynamically provisioned Workflow Managers (WfMs) manage workflows
    - Standing correlations could be requested that assess correlations as new data becomes available
      - Automatic notifications possible
  - CEs can federate to exchange information
  - Results ultimately displayed on CORE Google Earth (GE)

- CEs maintain a federated, distributed data service using iRODS
  - iRODS (integrated Rule-Oriented Data Service) is open source from University of California, San Diego and the Renaissance Computing Institute, North Carolina
  - iRODS agents can federate to enforce data policy, e.g.,
    - Automatic data replication to reduce latency & improve overall performance
    - Exchange of metadata to enhance discovery of correlations

- Integrated performance model enables correlation engines to estimate the computational load
  - CEs can federate to estimate total load on the DA and throttle accordingly
  - Give feedback to analyst on potentially intractable queries
Workflow Management (WfM) Engines

- Workflow Managers "orchestrate" or "choreograph" multiple steps in a large, distributed application
  - *Data Movement and Process Execution*

- Keeping track of which operations:
  - *Need to be done*
  - *Have completed*
  - *May have failed and need to be retried*

- Many Script-based or visual programming tools available to define workflows
  - *BPEL (Business Process Execution Language) most widely known in business community but problematic interoperability*
  - *Pegasus, Taverna, Triana, DAGMan widely known in science community*

- A workflow manager may only be needed here if correlation queries get extremely complicated
Functional View of Complete System

Database Aggregator

Private Clouds at Different Sites

Correlation Evaluator_1

WfM

Local Storage

iRODS

Correlation Engine w/ perf model

CORE_1
w/ CEC

Multiple CORE Users at Different Sites

Correlation Evaluator_n

WfM

Local Storage

iRODS

Correlation Engine w/ perf model

CORE_i
w/ CEC

CORE_j
w/ CEC

CORE_k
w/ CEC
End-to-End system with Google Earth-based clients using cloud-hosted, RulePoint Correlation Engines (CEs) to identify related events in multiple data sources.
Agent Logic’s RulePoint Basic Architecture

- Data Source
- Service
- Rules
- Event
- Responses

Arrows:
- poll
- subscribe
- subscribe
- results
Eucalyptus Basic System Architecture

- User (Euca2ools)
- Cloud Manager
- Cloud Controller
- Walrus Storage (VMI repository)
- Node 1
- Node 2
- Node n

Front-end

Back-end
Status of Initial Critical Tasks

1) Small private Eucalyptus cloud stood-up
   - *Eucalyptus is an open source, API clone of Amazon EC2*
   - *Project Penguin is a possible host platform*
   - *Perform "microbenchmarks" to evaluate performance/behavior*

2) Basic Correlation Evaluation (CE) basic package implemented
   - "*Packages": existing CORE concept for correlations against DA*
   - *Built CE virtual machine image (VMI) for dynamic provisioning in cloud*
   - *Initially running manually for development and testing*

3) Currently Modifying CORE to initiate CE VMIs in cloud
   - *Automating CORE "packages" for doing correlations*
   - *Add "Correlation Evaluation Client" window in CORE*
   - *Compliments CORE "tree"*
   - *Results from CEs in the cloud displayed on Google Earth*

4) Demonstrations and Evaluations to be done
   - *Perform benchmarks to evaluate performance/scalability/behavior*
   - *Add multiple CORE users*
   - *Add multiple client sites*
Further Enhancements

5) Integrated Performance Model & Control
   – Design performance model that estimates how much processing time and how much correlation data may be involved
   – Use Perf Model to prevent inadvertent initiation of "expensive" queries that result in excessive processing and data requirements
   – Use Perf model to "throttle" aggregate queries against DA so as to not adversely affect operational DA use

6) Integrated Workflow Management
   – CORE package may require extensive "query sweeps" against the DA
   – Use a Workflow Management engine to manage query sweeps

7) Integrated Data Management and Data Policy Enforcement
   – Use iRODS to manage correlation data across sites
   – Enforce data policy, e.g., caching, replication, security, transcoding

8) Development of Additional Cloud Applications
   – Demonstrate generic hosting capabilities
   – Multi-tenancy
   – Increased utilization
Potential Issues and Conclusions

• There are no cloud standards
  – *Amazon EC2 is de facto standard for IaaS*
  – *Work underway in OGF OCCI to standardize this interface*
  – *There is a risk that “standard” cloud APIs may be different*
  – *Critically important when federating clouds from different organizations*

• Cloud Architecture and Configuration
  – *CE application should be suitable for generic Eucalyptus configuration*
  – *Other apps may need specialized configurations for HPC, data streaming*
  – *How to manage cloud performance across NSS job mix?*

• More likely to realize benefits of Cloud Computing -- IaaS – when applications are run *at scale*
  – *Small demo application can be used to drive installation of cloud prototype but may not allow huge benefits to be demonstrated*
  – *More users/apps may be needed to show control of server utilization, etc.*
  – *May want to demonstrate workload management across sites*

• Green IT is another potential benefit of cloud computing
  – *Workflow mgmt across sites to enforce energy consumption policy*
Legal Notice

“All trademarks, service marks, and trade names are the property of their respective owners”