Facilitating self-adaptable Inter-Cloud management

G. Kecskemeti, M. Maurer, I. Brandic, A. Kertesz, Zs. Nemeth, S. Dustdar

20th Euromicro International Conference on Parallel, Distributed and Network-Based Processing
Feb 17, 2012.

https://www.lpds.sztaki.hu/CloudResearch
http://s-cube-network.eu
Intro to IaaS behavior

Infrastructure as a Service Cloud
Intro to federations

- Several public and private IaaS systems co-exist
  - Only a “Federated Cloud” could aggregate the different capabilities

- Federations can be defined on various levels
  - User oriented federations
    - Reducing the effects of provider lock-in
  - IaaS initiated federations
    - Users are not aware that they use a federated infrastructure

- Issues of federated infrastructures
  - Cross provider SLAs
  - Different appliance formats
  - Different APIs and UIs to access the cloud functionality
Federated Cloud Management (FCM)

- An autonomic resource management solution
- Provides an entry point to a cloud federation
- Provides transparent service execution for users
- Following challenges are considered:
  - Varying load of user requests
  - Enabling virtualized management of applications
  - Establishing interoperability and provider selection
  - Minimizing Cloud usage costs
- Builds on *meta-brokering, cloud brokering* and automated *on-demand service deployment*
- Layered architecture
  - Meta-broker
  - Cloud Brokers
  - Cloud infrastructure providers

FCM Architecture: overview

- **Top-level brokering**
- **Autonomously manage the interconnected cloud infrastructures**
- **Forms a federation with the help of Cloud Brokers**

---

FCM Architecture: overview

- Manages VA distribution among the various cloud infrastructures
- Automated federation-wide repository content management
- Offers current VA availability and estimates its future deployment
FCM Architecture: overview

- **Interacts** with a single IaaS system
- **Manages** resources
- **Schedules** service calls
The Cloud-Broker performs *scheduling* of service calls to resources (VMs)
- Based on the monitoring information gathered
- May *decide* to start new resources based on:
  - The number of running VM’s to handle the service call
  - The number of waiting service calls in the Service call queue
  - The average execution time of service calls
  - The average deployment time of VA’s
  - SLA constraints
- VM decommission
  - Takes into account the “billing period”
Autonomous behavior

• Inter-Cloud management for **optimized** resource usage and **SLA violation prevention**

• Predefined set of **reactive actions** in the Knowledge management system requiring local/global intervention in the system

<table>
<thead>
<tr>
<th>Action</th>
<th>Involved Component</th>
<th>Integration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reschedule calls</td>
<td>Meta-Broker</td>
<td>Global</td>
</tr>
<tr>
<td>Rearrange VM queues</td>
<td>Cloud-Broker</td>
<td>Global</td>
</tr>
<tr>
<td>Extend/Shrink VM Queue</td>
<td>Cloud-Broker</td>
<td>Local</td>
</tr>
<tr>
<td>Rearrange VA storage</td>
<td>FCM repository</td>
<td>Global</td>
</tr>
<tr>
<td>Self-Initiated Deployment</td>
<td>Service instances</td>
<td>Local</td>
</tr>
</tbody>
</table>

• Adaptation actions are triggered by a **rule-based system**, based on monitored metrics
• Knowledge manager can make fine-grained changes
  – involving actions on non-public interfaces
• Local reactive actions could cause an \textit{autonomic chain reaction}, where a single SLA violation prediction might lead to an unstable system
Global KM integration

- Makes architecture-wide decisions from an external viewpoint
  - Considers the state of the entire system before changing one of its subsystems
- Aggregates the metrics received from the different monitoring solutions
- Early adaptation action exhaustion
  - because of metrics aggregation and restricted interface use
Hybrid KM integration

- Hybrid approach for incorporating a Knowledge Management System to FCM
  - Combines global and local KM integration
- Allows global control over local decisions
  - Global KM could stop the application of a locally optimal action to avoid an autonomic chain reaction
  - Enables the execution of more fine-grained actions postponing adaptation action exhaustion
Adaptation actions

1 rule “Remove VA from native repository due to high local costs”

2 WHEN
   - C:3 r : NativeRepository()
   - In4 VA_x : VirtualAppliance()
   - std(locstcost(., VA_x))
   - deployfreq(r, VA_x) < mean(deployfreq(., VA_x))

3 THEN
   - In8 remove(VA_x, r) // removes VA_x from native repository

- Reschedule service calls
  - Cancel Ncr calls at the source cloud
  - Initiate rescheduling

- Rearrange VM queues
  - Migrate Nvmtr VMs from the source cloud to the destination

- Queue extension/shrinking
  - Increase/decrease the amount of VMs processing a particular service call

- Rearrange VA storage
  - Move VAs or parts of VAs from a particular repository to another one

- Self-initiated deployment
  - Upon local service overload/function loss, instantiate new VM
  - Create a proxy to forward calls to the new VM
Monitored metrics

- Service call queue length in every Cloud-Broker
- VM queue length for every appliance in every Cloud-Broker
- Call throughput
  \[ \text{throughput}(x) := \frac{1}{\max_{c \in Q_x} (\text{waitingtime}(c))} \]
- Average waiting time for particular service
  \[ \text{awt}(s, Q_x) := \frac{\sum_{c \in q(x,s)} \text{waitingtime}(c)}{|q(x, s)|} \]
- Average waiting time of a queue
- Number of service \((s)\) instances in an IaaS system \((x)\):
  \[ vms(x,s) \]
- Call/VM ratio
- overall infrastructure load
- Global storage cost
Conclusions

- We have designed a Federated Cloud Management solution that acts as an entry point to cloud federations
  - Meta-brokering, cloud brokering and on-demand service deployment
- We have extended the FCM architecture with autonomous behavior
  - Using a hybrid knowledge management system and rule based autonomous manager
- Future works
  - Autonomy should also consider green aspects
  - Alternative knowledge management systems
  - Performance measurements on a simulated and on a physical system that has the autonomous manager(s) enabled
Thank you for your attention!

Questions?

For more details have a look at the webpage of our cloud research group at MTA SZTAKI LPDS:

https://www.lpds.sztaki.hu/CloudResearch