CloneCloud: Elastic Execution between Mobile Device and Cloud

Laura Vasiliu

21st March, 2012
Motivation

Mobile cloud applications
Motivation

Strong connectivity

Mobile cloud applications
Motivation

Strong connectivity

computing resources on commercial clouds

Mobile cloud applications
Motivation

Mobile cloud applications

- Strong connectivity
- Computing resources on commercial clouds
- Storage resources on commercial clouds
Motivation

Mobile cloud applications

- Strong connectivity
- Computing resources on commercial clouds
- Storage resources on commercial clouds
- Faster execution time
Motivation

Mobile cloud applications

- Strong connectivity
- Computing resources on commercial clouds
- Storage resources on commercial clouds
- Lower energy use
- Faster execution time
Overview

- What is CloneCloud?
- Application partitioning
- Static analyzer
- Dynamic profiler
- Optimization solver
- Distributed execution
- Evaluation
- Conclusions
Overview

- What is CloneCloud?
  - Application partitioning
  - Static analyzer
  - Dynamic profiler
  - Optimization solver
  - Distributed execution
  - Evaluation
  - Conclusions
What is CloneCloud?

- System model that automatically transforms unmodified mobile application by off-loading the right portion of their execution onto device clones
What is CloneCloud?

- flexible application partitioner that can adapt to different environments
- for applications targeting the application VM
- optimizes execution time and energy use
- programmer must not concern about the application partitioning
Overview

- What is CloneCloud?
- Application partitioning
- Static analyzer
- Dynamic profiler
- Optimization solver
- Distributed execution
- Evaluation
- Conclusions
Application partitioning 1/2

Partitioning analysis framework
class C {
    void a () {
        if () {b(); c();}
    }
    void b() {
        // lightweight
    }
    void c() {
        // expensive
    }
}
void main () {
    C c; c.a();
}

(a) program

(b) static control-flow graph

(c) partitioned graph

- System model picks which parts of an application’s execution to retain on the mobile device and which to migrate to the cloud
Overview

- What is CloneCloud?
- Application partitioning
- **Static analyzer**
- Dynamic profiler
- Optimization solver
- Distributed execution
- Evaluation
- Conclusions
identifies legal choices for placing migration and re-integration points in the code, according to a set of constraints

**Constraints**

1. Methods that access specific features of a machine must be pinned to the machine
2. Methods that share native state must be collocated at the same machine
3. Prevent nested migration
Overview

- What is CloneCloud?
- Application partitioning
- Static analyzer
- **Dynamic profiler**
- Optimization solver
- Distributed execution
- Evaluation
- Conclusions
Dynamic Profiler

➢ profiles the input executable on different platforms and returns a set of profiled executions used to compose a cost model

**Input**: randomly chosen set of data

**Output**: set of profiled executions, each with 2 profile trees
Profile tree = representation of an execution on a single platform
Dynamic Profiler

Cost models

- execution time
  - collect timings at method entry and exit points

- energy consumption
  - <CPU activity; Display state; Network state> power level function
Overview

- What is CloneCloud?
- Application partitioning
- Static analyzer
- Dynamic profiler
- **Optimization solver**
- Distributed execution
- Evaluation
- Conclusions
Optimization Solver

- Chooses from the legal partitions which methods to migrate to the clone for minimizing the costs
- Places the migration points
- Modifies the application binary

Output: value assignment to binary decision variables $R(m)$
- $R(m) = 1 \Rightarrow$ place a migration point
- $R(m) = 0 \Rightarrow$ unmodified in application binary
Optimization Solver

- Goal: minimize the cost of a partition

\[ \text{Partition\_cost}\ (E) = \text{Computation\_cost}(E) + \text{Migration\_cost}(E) \]
Overview

- What is CloneCloud?
- Application partitioning
- Static analyzer
- Dynamic profiler
- Optimization solver
- Distributed execution
- Evaluation
- Conclusions
Distributed Execution

- implement a specific partition of an application process running inside an application-layer virtual machine
- operates at the granularity of a thread
- happens at runtime
Migration components

CloneCloud: Elastic Execution between Mobile Device and Cloud
Laura Vasiliu
How migration works

Execution of the process reaches migration point
How migration works

Execution of the process reaches migration point

Executing thread is suspended, its state is packaged and shipped to a synchronized clone
How migration works

Execution of the process reaches migration point

Executing thread is suspended, its state is packaged and shipped to a synchronized clone

Thread state is instantiated into a new thread with the same stack and reachable heap objects and then resumed
How migration works

Execution of the process reaches migration point

Executing thread is suspended, its state is packaged and shipped to a synchronized clone

Thread state is instantiated into a new thread with the same stack and reachable heap objects and then resumed

Migrated thread reaches reintegration point, its state is packaged and shipped back to the mobile device
How migration works

Execution of the process reaches migration point

Executing thread is **suspended**, its state is packaged and shipped to a synchronized clone

Thread state is instantiated into a new thread with the same stack and reachable heap objects and then **resumed**

Packaged thread is **merged** into the original process

Migrated thread reaches reintegration point, its state is packaged and shipped back to the mobile device
How migration works

- map objects at the original address space to the objects they “became” at the cloned address space
- MID = Mobile ID
- CID = Clone ID

Object mapping table
Overview

- What is CloneCloud?
- Application partitioning
- Static analyzer
- Dynamic profiler
- Optimization solver
- Distributed execution

Evaluation

Conclusions
### Evaluation 1/4

**Tested applications**
- Virus scanner (VS)
- Image search (IS)
- Behavior profiling (BP)

<table>
<thead>
<tr>
<th>Application</th>
<th>Input Size</th>
<th>Phone Exec.(sec)</th>
<th>Clone Exec.(sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VS</td>
<td>100KB</td>
<td>6.1</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>1MB</td>
<td>59.3</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td>10MB</td>
<td>579.5</td>
<td>22.5</td>
</tr>
<tr>
<td>IS</td>
<td>1 img</td>
<td>22.1</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>10 img</td>
<td>212.8</td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td>100 img</td>
<td>2122.1</td>
<td>79.2</td>
</tr>
<tr>
<td>BP</td>
<td>Depth 3</td>
<td>3.3</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>Depth 4</td>
<td>52.1</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>Depth 5</td>
<td>302.7</td>
<td>10.9</td>
</tr>
</tbody>
</table>
## Evaluation 1/4

### Tested applications
- Virus scanner (VS)
- Image search (IS)
- Behavior profiling (BP)

<table>
<thead>
<tr>
<th>Application</th>
<th>Input Size</th>
<th>Phone Exec.(sec)</th>
<th>Clone Exec.(sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VS</td>
<td>100KB</td>
<td>6.1</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>1MB</td>
<td>59.3</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td>10MB</td>
<td>579.5</td>
<td>22.5</td>
</tr>
<tr>
<td>IS</td>
<td>1 img</td>
<td>22.1</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>10 img</td>
<td>212.8</td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td>100 img</td>
<td>2122.1</td>
<td>79.2</td>
</tr>
<tr>
<td>BP</td>
<td>Depth 3</td>
<td>3.3</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>Depth 4</td>
<td>52.1</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>Depth 5</td>
<td>302.7</td>
<td>10.9</td>
</tr>
</tbody>
</table>
Migration costs

- WiFi latency of 69ms and bandwidth of 6.6Mbps
- 3G had latency of 680ms and bandwidth of 0.4Mbps
- include the cost of merging the thread
- network-specific transmission of the thread state
Evaluation 3/4

CloneCloud: Elastic Execution between Mobile Device and Cloud

Laura Vasiliu

25/29
Observations

- larger workloads benefit from off-loading more
  - due to amortization of the migration cost over a larger computation at the clone that receives a significant speedup

- energy consumption mostly follows execution time
  - energy expenditure is proportional to how long it is waiting for a response
Overview

- What is CloneCloud?
- Application partitioning
- Static analyzer
- Dynamic profiler
- Optimization solver
- Distributed execution
- Evaluation
- Conclusions
Conclusions

- Up to 20x speedup
- 20x energy reduction
- Flexible architecture
- No work for the applications developer
CloneCloud makes applications fast and energy efficient without programmers involvement!