Container Start Times: Empirical Analysis and Predictability

12th Symposium on Software Performance

Martin Straesser, Samuel Kounev

10.11.2021

Leipzig, Germany

https://se.informatik.uni-wuerzburg.de
Container start times are interesting for various domains

- How to build containerized software with fast start times?
- System simulation
- Evaluation of container engines
- Optimized deployments
- Resilience
- Scaling
- Serverless Computing

Motivation
Problem Statement

- What is the start time of a container?
  - The time difference between the following two events
    - Event 1: The container engine (e.g. Docker) receives the command to start a new container
    - Event 2: The container starts the execution of the `entrypoint` command
  - Note: This time difference can be evaluated without deeper knowledge about the container and without dependencies of the container (in contrast to readiness time)

- Central research questions
  - What are influencing factors for container start times?
  - Can we predict the start time of an arbitrary container?

- Idea: Test a large set of containers from the public repository Docker Hub
Influencing Factors For Container Start Times

**Image**
- Image size
- Number of layers
- Size of layers
- Platform architecture
- Platform OS
- Number of exposed ports
- Volumes
- Start command
- ...

**Deployment**
- (Image in local repo?)
- Node architecture
- Node OS
- Node kernel
- Container runtime
- Storage driver
- Number of containers deployed
- ...

Container Start Times: Empirical Analysis and Predictability

*Martin Straesser*
## Study Phases

<table>
<thead>
<tr>
<th>Design and Development</th>
<th>Measurements in Private Cloud</th>
<th>Measurements in Public Cloud</th>
<th>Result Analysis</th>
</tr>
</thead>
</table>
| • Setting the project scope  
• Study and setup design  
• Software development | • Process and software testing  
• Determine settings for public cloud (e.g., measurement repetitions, multi-threading, …) | • Container start time measurements in a representative environment | • Data analysis and visualization  
• Prediction model  
• Identify open questions and possible future work |
Experiment Setup

Cloud VM

Measurement Process

Scheduler

Image and Management DB

Monitoring

Collectors

Result DB
## Result Data and Contributions

### Image Metadata

<table>
<thead>
<tr>
<th>ImageID</th>
<th>Image Size</th>
<th>Volumes</th>
<th>Ports</th>
<th>Layers</th>
<th>…</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>12</td>
<td>5</td>
<td>2</td>
<td>32</td>
<td>…</td>
</tr>
<tr>
<td>y</td>
<td>45</td>
<td>1</td>
<td>1</td>
<td>13</td>
<td>…</td>
</tr>
<tr>
<td>…</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>z</td>
<td>34</td>
<td>0</td>
<td>1</td>
<td>8</td>
<td>…</td>
</tr>
</tbody>
</table>

### Measurement Results

<table>
<thead>
<tr>
<th>ImageID</th>
<th>Start Time</th>
<th>…</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>1231</td>
<td>…</td>
</tr>
<tr>
<td>x</td>
<td>1564</td>
<td>…</td>
</tr>
<tr>
<td>…</td>
<td></td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>2344</td>
<td>…</td>
</tr>
<tr>
<td>y</td>
<td>2654</td>
<td>…</td>
</tr>
<tr>
<td>…</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Contributions**
  - Dataset
  - Statistical measures starting one specific image
  - Regression data (manifest parameters → start time)
  - Prediction model
  - Others (e.g. evolution of start times over different versions)
Extension Points and Recent Challenges

Extension points
- Validating results for start times in environments with container orchestration (e.g. Kubernetes)
- Extend evaluation with parameters from deployment (e.g. host OS, storage driver, …)
- Evaluate other container engines
- …

Recent challenges
- Getting enough pullable images from Docker Hub (no full list available and search results are always limited to 400 images)
- Developing the software for measurement execution and result collection (databases, scheduling, live monitoring, failure prevention and handling, …)
## Methodology

- We explicitly consider best practices from latest ACM SIGSOFT Empirical Standards [1]
- We maintain a question catalogue which answers fundamental questions for all stages of the study

### Table: Methodological Approaches

<table>
<thead>
<tr>
<th>Category</th>
<th>Approaches</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Standard</td>
<td>Action Research, Benchmarking, Case Study</td>
</tr>
<tr>
<td>Data Science</td>
<td>Engineering Research, Experiments, Grounded Theory</td>
</tr>
<tr>
<td>Meta Science</td>
<td>Mixed Methods, Optimization Studies, Qualitative Surveys</td>
</tr>
<tr>
<td>Questionnaire Surveys</td>
<td>Systematic Reviews, Longitudinal, Quantitative Simulation</td>
</tr>
</tbody>
</table>

Methodology

1. Preliminaries
   1. Motivation
   2. Problems, objectives, research questions
   3. Related key concepts
   4. Relationship to other studies

2. Study Execution
   1. Overview of methodology
   2. Data description
   3. Environment of data acquisition
   4. Temporal aspects of data acquisition
   5. Data collection and processing

3. Investigation of Results
   1. Data analysis
   2. Result presentation
   3. Answers to research questions

4. Wrapping up
   1. Assumptions, limitations and threats to validity
   2. Replicability
   3. Open questions and future work

5. Supplementary Materials
   1. Datasets and source code
   2. Technical documentation
Conclusion

Problem

• No systematic study on container start times with a large number of images in a representative environment
• Influencing factors are only partially known

Idea

• Measure the start time of a large set of containers from a public repository
• Analyze image and deployment factors

Benefits

• Deeper knowledge about container start times is beneficial in various domains, e.g., development of containerized software, serverless computing, …