Towards Predicting Resource Demands and Performance of Distributed Cloud Services

Sevil Dräxler, Manuel Peuster, Marvin Illian and Holger Karl
Overview

- Motivation
- Experimental Setup
- Data Analysis
- Conclusion and Discussion
Imagine you are a network service provider …

You want to know about a service’s behavior prior to its deployment!
Typical Network Service Orchestration

Developer

S → A → B → C → D

Placement

Scaling

Monitoring

Operator

S. Dräxler, M. Peuster, M. Illian, H. Karl, Towards Predicting Resource Demands and Performance of Distributed Cloud Services
Problem: Limited Precision and Flexibility

- Network services and functions specified using descriptors...

```yaml
image: ubuntu-14.04-server-amd64-disk1
scale_in_out: 2
disk_size: 10 GB
mem_size: 2048 MB
num_cpus: 2
```

- Service developer should know or estimate:
  - Exact resource demands of service components
  - Number of required service component instances

→ Risk of over-/under-estimating resources
Overview

- Motivation
- *Experimental Setup*
- Data Analysis
- Conclusion and Discussion
Testbed

- OpenStack testbed for measurements
- 1 controller node
- 3 compute nodes
Source videos

- 5 different videos (same duration)
- Resolution: 1920x1080 ("Full HD")
- Frame rate: 25-60 frame/s
- Playback data rate: 3.5-14 Mbps
- Different characteristics
  - Colorful/black & white, fast/slow pacing, grainy picture, ...
Video encoding VNF

- **Parameters**
  - Source video files
  - Target resolutions
  - Target video playback data rates
  - Frame rates
  - Assigned CPU cores

- **Models**
  - Min. required CPU cores
  - Min. required memory
  - Achievable framerate
Cache VNF

- **Parameter**
  - Input data rate (determined by data rate of video encoding VNF)

- **Models**
  - CPU utilization
  - Memory utilization
Overview

- Motivation
- Experimental Setup
- *Data Analysis*
- Conclusion and Discussion
Model Extraction

- Support vector regression
  - Supervised machine learning
  - Find flat function with small error

- Polynomial regression
  - Minimize sum of squared deviations between predictions and observed values
  - Varying degrees
  - Visual inspection to avoid under-/over-fitting

- Metric for deviation from training data
  - Mean squared error:
    \[
    \frac{1}{n} \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2
    \]
Model: Required CPU cores (video encoder)

Training data: Resolution, bitrate, framerate, min. vCPU

SVR (MSE=1.55)  
PR (MSE=1.8)
Model: CPU utilisation (cache)

Training data: Resolution, encoder vCPUs, CPU utilization

SVR (MSE=2.7)  PR (MSE=0.7)
Non-trivial proportionalities

CPU Utilization of Cache VNF

Target encoding datarate (Kb/s)

CPU Utilization (%)

Encoder VNF
- 1 CPU
- 2 CPUs
- 3 CPUs
- 4 CPUs

S. Dräxler, M. Peuster, M. Illian, H. Karl, Towards Predicting Resource Demands and Performance of Distributed Cloud Services
Overview

- Motivation
- Experimental Setup
- Data Analysis
- Conclusion and Discussion
Conclusion

- Characterizing resource demands and performance metric values of VNFs is feasible

Service component

Resource demands

- CPU: $3\cdot \lambda_1 + \lambda_2 + 1$
- Mem: $\lambda_1 + 2\cdot \lambda_2 + 5$

Input data rates

- $\lambda_1$
- $\lambda_2$

Output data rates

- $2\cdot \lambda_1$
- $0.5\cdot \lambda_2$
Conclusion

- Characterizing resource demands and performance metric values of VNFs is feasible
  - Can be used for efficient and flexible placement and scaling of services

Conclusion

• Characterizing resource demands and performance metric values of VNFs is feasible
• VNFs have to be profiled in their final SFC setup
Conclusion

- Characterizing resource demands and performance metric values of VNFs is feasible
- VNFs have to be profiled in their final SFC setup
  - Challenges: test specification, automatic profiling of SFCs

References

- Our experimental is data available online
  - https://uni-paderborn.sciebo.de/s/G9q2hmUNg4n8LEg
