Improving Decision-Making for Nearby Compute Deployment

Mobile and Pervasive Computing (Fall 2022)

Students: Patrick Coppock, Sofia Martins

Mentor: Jan Harkes

Background

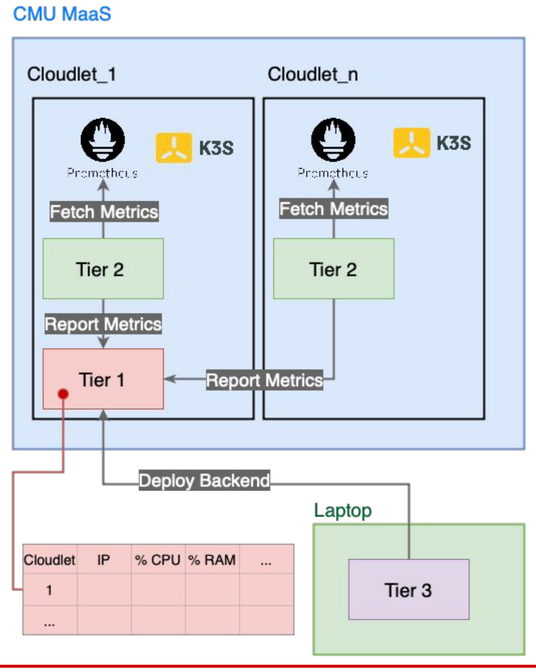
- ***Sinfonia***[1]*:* a system for mobile applications to discover and deploy compute intensive operations on nearby infrastructure
- Sinfonia finds and dynamically associates an app launched on a Tier-3 device with its software back-end on a Tier-2 cloudlet

Problem Statement

- Sinfonia's cloudlet selection algorithm is simplistic, as it is location-based
- **Solution:**
  - Bring in additional Tier-2 metrics to enrich cloudlet allocation algorithms
  - Implement other task allocation algorithms (Match Functions)

Architecture

- Cloudlets are deployed in the form of Kubernetes clusters
- Prometheus instance running within each cloudlet to fetch resource metrics (CPU, memory, disk, network, GPU)
- Cloudlets deployed on a MAAS cluster (maas.cmusatyalab.org)
- Two compatible Tier-3 applications:
  - Bash
  - Open Real-Time Style Transfer (OpenRTiST[2])



New Metrics

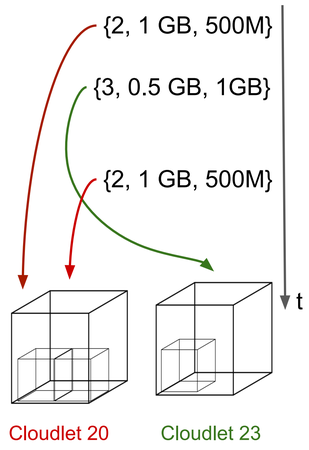
- Originally, only relative capacity (%) was being reported for each cloudlet resource (CPU, memory, …) - Relative [R] metrics
- To augment Sinfonia's decision making algorithms, we introduce Absolute [A] metrics

Metrics that were originally implemented

New metrics

	Usage Metrics
CPU	[R] Percentage Used [A] Time spent in non-idle modes (s)
RAM	[R] Percentage Used [A] Free Memory (B)
GPU	[R] Percentage Used
NIC	[A] Rx Throughput [A] Tx Throughput
Disk	[A] Free Disk (B)

New Match Functions



- Formulate the task placement problem as a Combinatorial Optimization Problem
- Tier-3 tasks are represented as items with given volume {CPU, Mem, Disk} that are to be placed into bins (cloudlets) of finite capacity
- **First-Fit** and **Best-Fit** Online Bin Packing functions are introduced
- A third function picks cloudlets so as to minimize the **L^2-norm of CPU and memory use** if the task is placed in it
- To reduce the action space: leverage the Power-of-two choices

Contributions

- New Tier-2 usage metrics were defined
- Three new matching functions were defined
- Web Dashboard integration using Grafana
- Performance Evaluation pipeline was introduced

Limitations

- Match function based on latency is still needed
- RTT estimation is difficult to achieve
- A Tier-3 could be behind NAT
- Root privileges needed to run ICMP
- Simple timing of an API endpoint might be more accurate anyway.

Future Work

- Tier 2-3 RTT measurements
- Demonstrate VM compatibility within Sinfonia
- Benchmarks for accurate characterization of client's resource requirements

References

1. M. Satyanarayanan *et al.*, 'Sinfonia: Cross-Tier Orchestration for Edge-Native Applications'
2. S. George *et al.*, 'OpenRTiST: End-to-End Benchmarking for Edge Computing', *IEEE Pervasive Computing*, vol. 19, no. 4, pp. 10–18, Oct. 2020, doi: 10.1109/MPRV.2020.3028781
3. B. Balaji, C. Kakovitch, and B. Narayanaswamy, 'FirePlace: Placing Firecracker Virtual Machines with Hindsight Imitation', in *Proceedings of Machine Learning and Systems*, 2021, vol. 3, pp. 652–663.