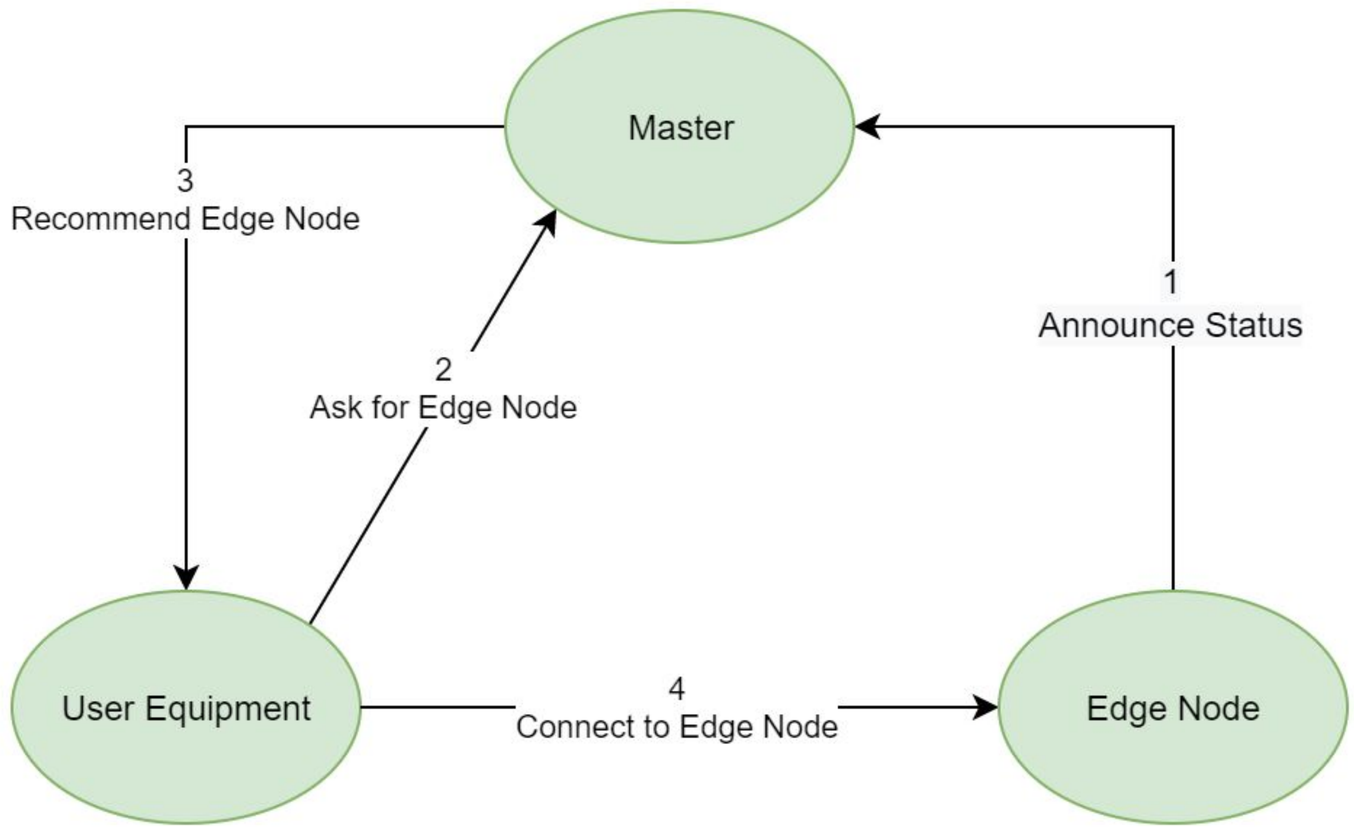
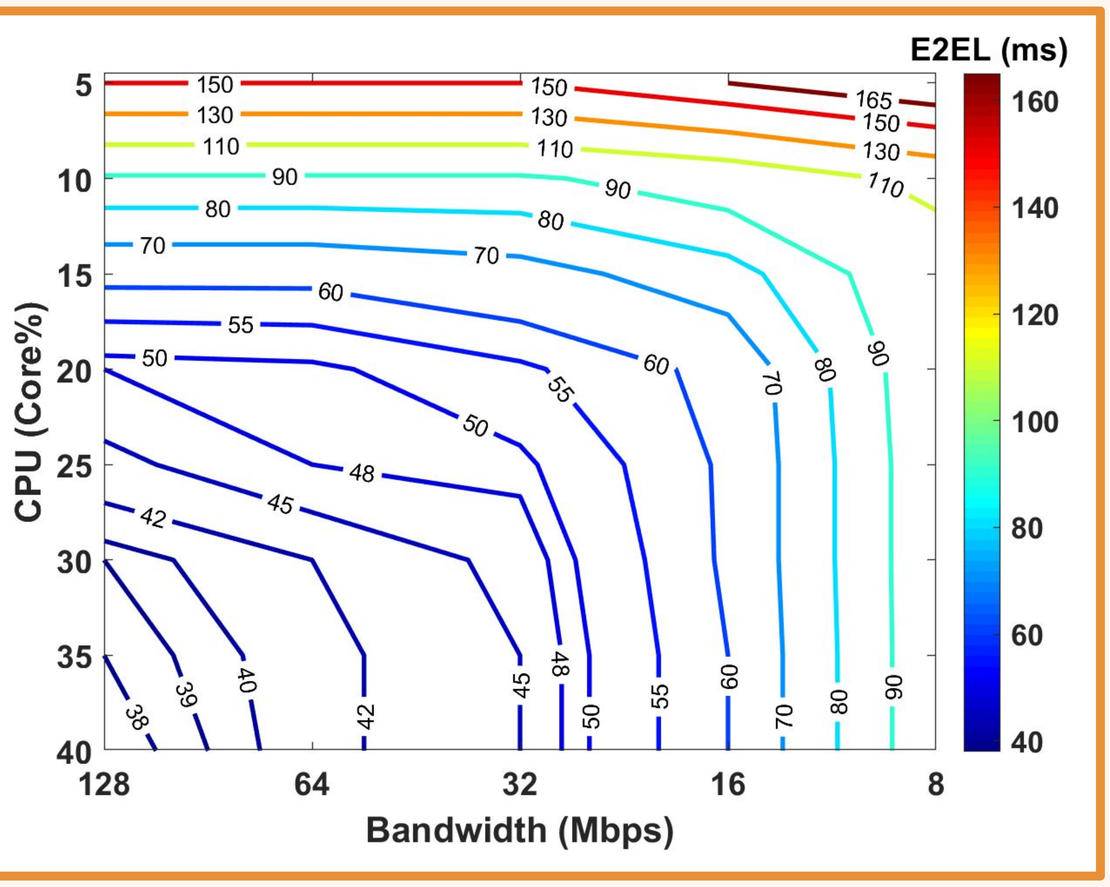
Edge Resource Allocation Based on End-to-End Latency

Aly Sakr, Rolf Schuster
Dortmund University of Applied Sciences and Arts
aly.sakr@fh-dortmund.de, rolf.schuster@fh-dortmund.de

Introduction:

Edge computing has a clear advantage over cloud computing when it comes to delay-sensitive applications. However, it also introduces new challenges, one of which is how to allocate edge resources to user equipment applications. That is how to determine that a particular edge resource and the connection to it is good enough for an application. We develop a practical approach based on End-to-End Latency (E2EL) as measured by the application itself. The approach is based on the model shown here, which aims to ensure an acceptable level of quality of service by assigning the responsibility of evaluating the connection to the user equipment.





End-to-End Latency:

E2EL as measured by user equipment is an accurate indicator for perceived quality of service. To support this argument we collected E2EL data from a real-world application (pedestrian tracking) under different conditions. We try to capture common scenarios: hardware congestion presented as CPU limit, and network congestion presented as bandwidth limit. Normally, such scenarios result in low quality of service, the aim of this experiment is to show that E2EL is also affected in these scenarios, and thus is a good indicator of quality of service. The results shown here demonstrate the effect of congestion (hardware and network) on E2EL, the lower the available resources are, the higher the E2EL is. The results also show a bottleneck behavior, that is, the lowest available resource determines the final E2EL.

Results:

The experiment conducted here shows the behavior of our model, when a point of access change results in an unacceptable quality of service. The user equipment successfully detects the situation and switches to local execution until a connection to another more suitable edge node is ready. The implementation works as follows: whenever E2EL is below a defined threshold (acceptable quality of service) the connection score is incremented, otherwise it is decremented. If the connection score hits a maximum value the user equipment stops local execution and switches to remote execution only. On the other hand if it hits zero the connection is terminated. Values in between indicates evaluation phase, where the connection is not terminated and local execution is still on.