Location-aware Load Prediction in Edge Data Centers

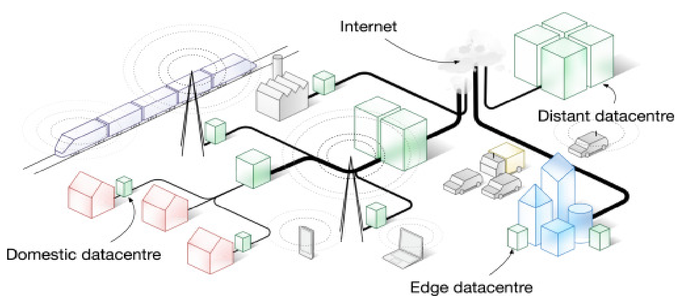
Chanh Nguyen, Cristian Klein, Erik Elmroth
chanh@cs.umu.se, cklein@cs.umu.se, elmroth@cs.umu.se

Umeå University
WASP WALLENBERG AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Description

We simulate Mobile Cloud Network as a hexagonal grid distributed across a geographical area, in which each node is supposed to connect and provide service to user in its vicinity. A **Location-aware Load prediction** algorithm based on the **Vector Autoregressive Model** is developed to estimate short term workload in every node. We utilize the real world mobility trace to simulate load in each node and conduct various experiments to evaluate the proposed algorithm. Result shows that our proposed algorithm is able to achieve an average accuracy of up to 93%, which slightly **improves predictions by 4.3%** compared to the state-of-the-art location-unaware prediction method.

Background & Motivation



Mobile Cloud Network (MCN) is a complemented platform to traditional centralized cloud [2] in which IT capabilities are moved closer to users in order to cut down on application-level latency. **Edge Data Center (EDC)** are attached to cellular base station or wireless access point and provide service to user in its proximity. The bounded coverage area of base station and limited capacity of each EDC in addition to the uncertainty of user location are challenging the resource management operator in capacity adjustment and planning. The problem can be solved with a proactive dynamic resource provisioning in which resource usage in each EDC is estimated in advance, which is made available for the decision making to efficiently determine various management actions (power on/off server, allocate/release extra resource for peak/idle periods, etc.), to ensure that EDC **persistently meets the SLAs of applications hosted, while maximizing resource utilization**.

Methods

Suppose $node_k$ has n neighbors, each with historical time series y_{it}. We apply the vector autoregressive model [5] to build the predictive model for $node_k$ which has the VAR(p) form:

$$Y_t = c + \Pi_1 Y_{t-1} + \Pi_2 Y_{t-2} + ... + \Pi_p Y_{t-p} + \epsilon_t, t = 1, ..., T$$

where $Y_t = (y_{1t}, ...y_{(n+1)t})'$ denotes $((n+1) \times 1)$ vector of time series variables formed by combining every single time series of $node_k$ and its n neighbors. Π_i are $((n+1) \times (n+1))$ coefficient matrices and ϵ_t is an $((n+1)) \times 1)$ unobservable zero mean white noise vector process with time invariant covariance matrix Σ.

Preliminary Results

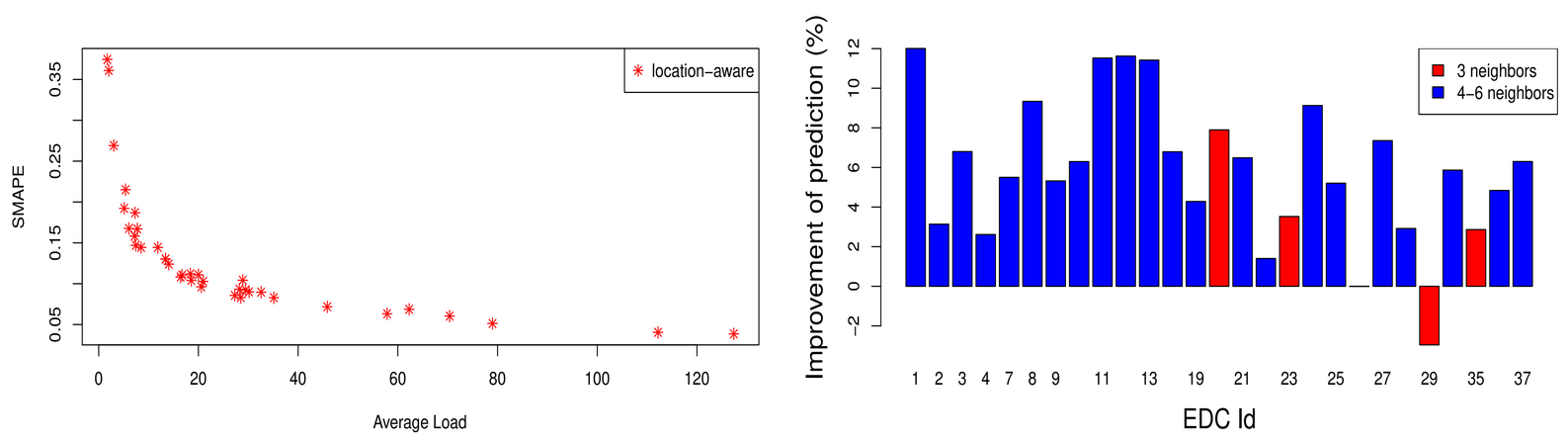


Figure 2: Left: The influence of average load on the prediction accuracy; Right: The improvement of predictions between two models in nodes with different neighbors.

We use the real mobility traces of taxis in San Francisco [4] to simulate the historical workload in the Edge Data Centers. The proposed method can return the average prediction accuracy from **64% to 96%** respected to the error metric SMAPE.

- The **high prediction** results occurred on those nodes with **substantial load**.
- Compared to the location-unaware prediction method using ARIMA [3], the proposed algorithm **outperforms with prediction accuracy improved by 4.3%**.
- The difference in prediction accuracy is not significant for nodes with 3 neighbors. However, the variation is considerable for nodes with **at least 4 neighbors**.

Roadmap & Milestones

- Further evaluation of the proposed algorithm with different practical use-cases.
- Work on devising a distributed prediction algorithm that runs on each edge node.

Research Goal & Question

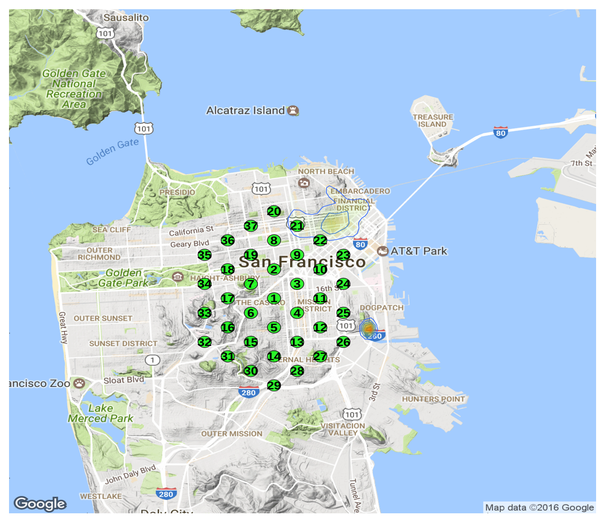


Figure 1: Example of hexagonal grid map.

Can workload predictions be improved by using knowledge about the neighborhood relationship of EDCs? Existing approaches do not take advantage of the potential correlation of load among nodes.

The goal of our research is first analyzing the impact of user mobility behavior on the resource usage at EDCs. We investigate the cross-correlation in workload changing among those EDCs located close each other when user location changed. Afterward, the final goal is to develop an efficient load prediction method which is capable of capturing and utilizing such knowledge about neighbor relationship of EDCs to improve the performance in term of prediction accuracy.

Bibliography

[1] William Tärneberg, Amardeep Mehta, Eddie Wadbro, Johan Tordsson, Johan Eker, Maria Kihl, Erik Elmroth. Dynamic application placement in the Mobile Cloud Network. Future Generation Computer Systems, 2016.

[2] Hu, Yun Chao, et al. "Mobile Edge Computing-A Key Technology Towards 5G." ETSI White Paper 11 (2015).

[3] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, Time series analysis: forecasting and control. John Wiley & Sons, 2015.

[4] M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser, "CRAWDAD dataset EPFL/mobility (v. 2009-02-24)", Downloaded from http://crawdad.org/epfl/mobility/20090224, Feb. 2009.

[5] Lütkepohl, Helmut. Vector autoregressive models. Springer Berlin Heidelberg, 2011.