ELASTIC: Edge Workload Forecasting based on Collaborative Cloud-Edge Deep Learning

Yanan Li[1], Haitao Yuan[2], Zhe Fu[3], Xiao Ma[1],

Mengwei Xu[1], and Shangguang Wang[1]

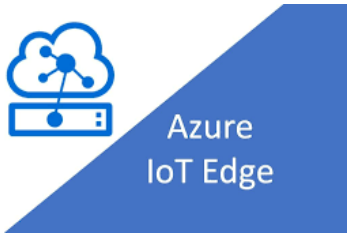
北京郵電大學 Beijing University of Posts and Telecommunications

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

清華大學 Tsinghua University

Background-Edge is Gaining Momentum

- In industry, many companies have launched edge computing services.



- More and more web services are adopting edge computing due to the advantages of **low latency** and **high bandwidth**.

More limited and lower resource utilization

of physical servers:

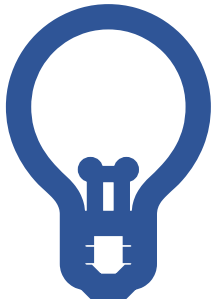
- Centralized cloud: thousands or even millions;
- Each edge site: tens or hundreds.

CPU utilization [1]:

- Centralized cloud: 47% VMs less than 10%;
- Edge platform: 74% VMs less than 10%.

[1] Xu, Mengwei, et al. "From cloud to edge: a first look at public edge platforms." *IMC*. 2021.

For edge service providers (ESPs), how to make the most of limited edge resources?



Edge workload forecasting

Workload forecasting is the cornerstone

- Usage:
  - Resource provisioning;
  - Request offload scheduling;
  - Adaptive bitrate (ABR) schemes.
- Effect:
  - [NSDI'21] Reduce operating costs up to 65%;
  - [WWW'21] Improve user experience more than 30%.

[2] Singh, Rachee, et al. "Cost-effective cloud edge traffic engineering with cascara." *NSDI*. 2021.
[3] Ye, Fanghua, et al. "Outlier-resilient web service QoS prediction." *WWW.* 2021.

Existing Forecasting Methods

1. “Edge-only” methods. Each edge site deploys their model separately and individually to only capture the correlation among VMs of same edge sites, namely intra-site correlations.
   Limitations: it does not (or poorly) consider the inter-site correlations.

2. “Cloud-only” methods. Each edge site transmit the whole workload data to the centralized cloud for centralized model training and inference.
   Limitations:
   1. Huge data transmission overhead;
   2. Additional time consumption of large models’ training and inferencing;

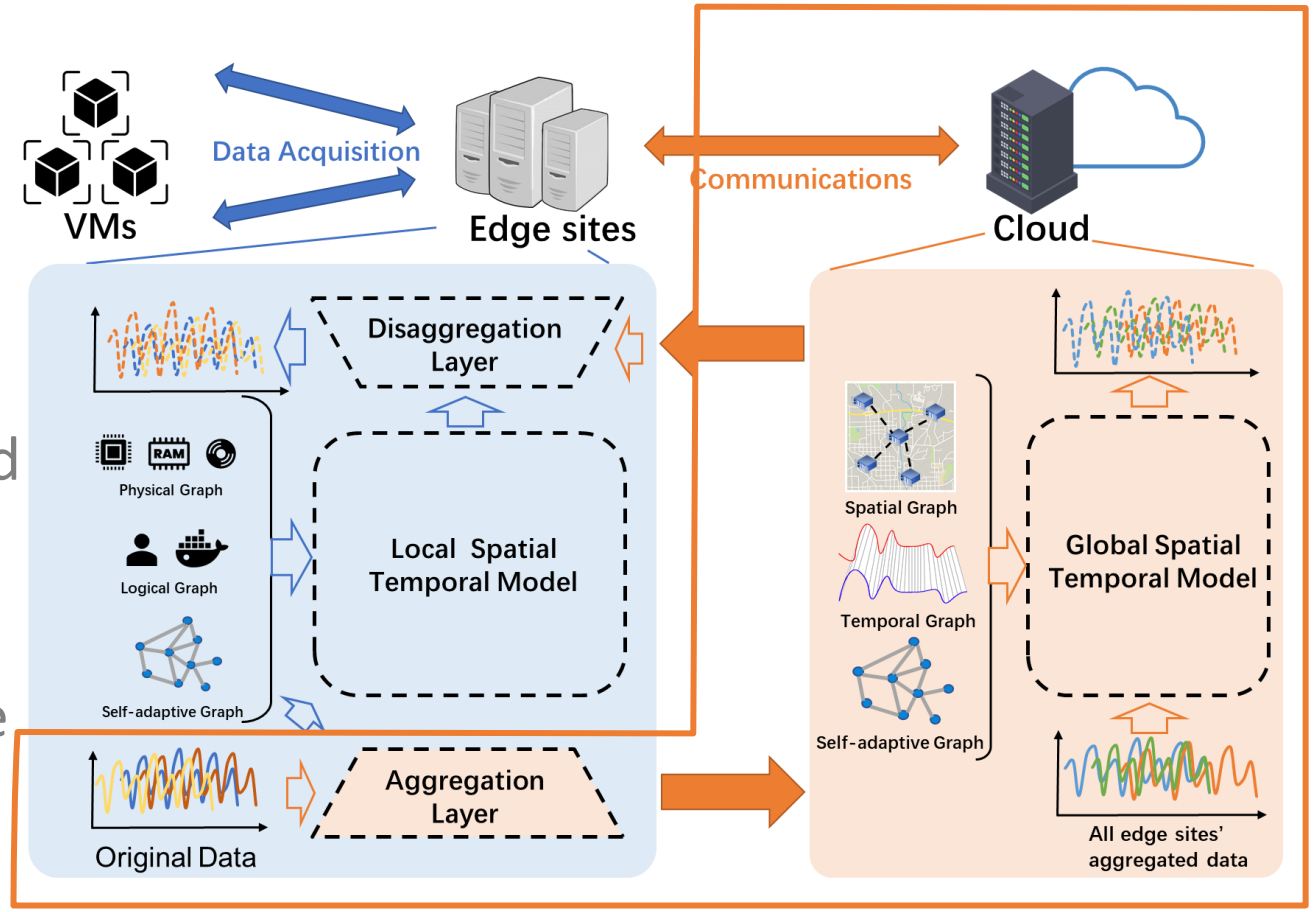
Existing Collaboration Methods

3. DNN model partition divides the giant model into different sub-models and deploys them on the cloud and the edge separately.
   Limitations: it is mainly applied to model inference and need an offline profiling phase for each edge site and cloud pair separately.
4. Federated learning is another type of collaboration paradigm to training DNN models on data distributed participants.
   Limitations: it requires all participants to train a common model, which cannot be directly applied to workload forecasting scenario, since each edge site generally has a different number of VMs.

ELASTIC: A Collaborative Cloud-Edge Approach

It mainly consists of two stages:

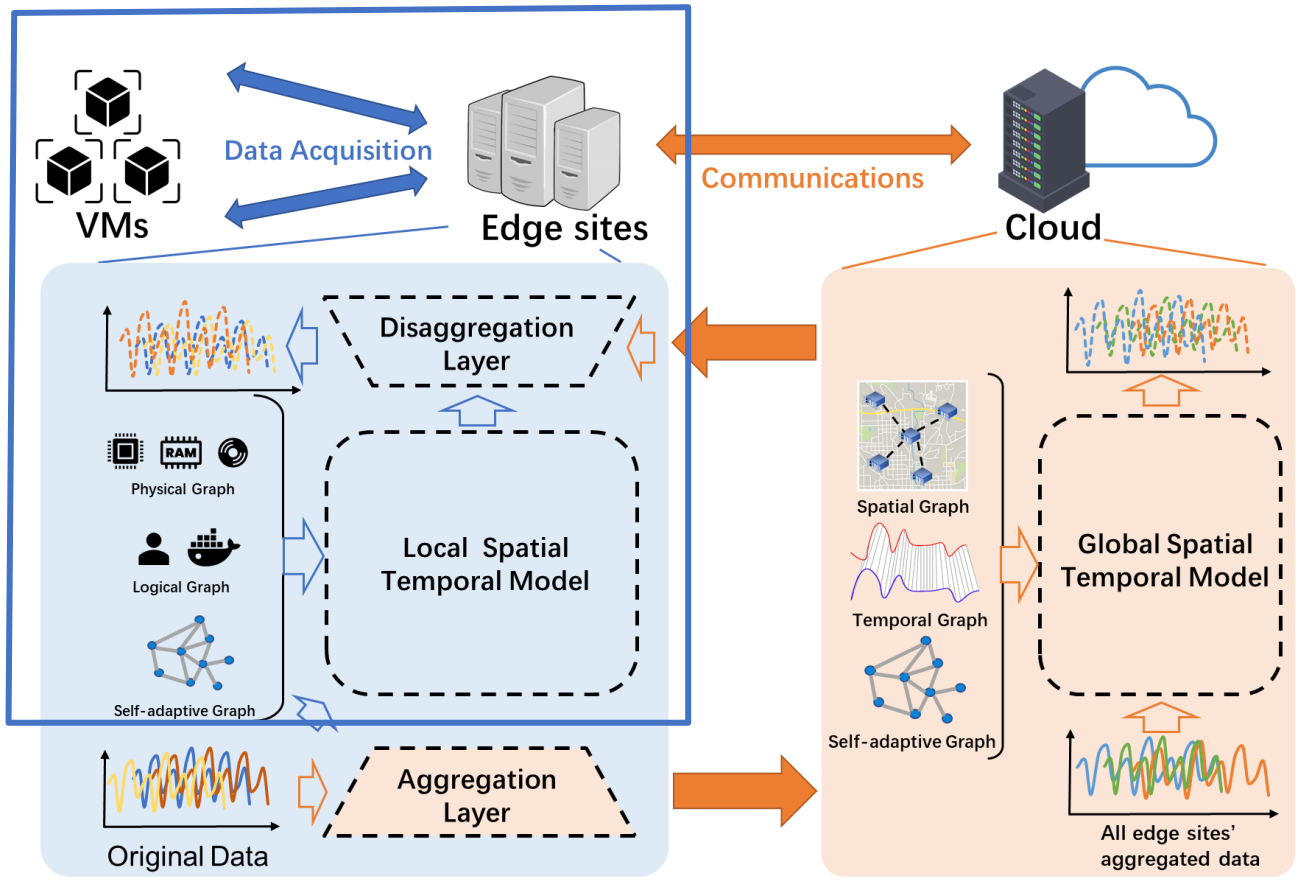
- The global stage performs coarse-grained spatial-temporal forecasting at the centralized cloud;
- The local stage performs fine-grained spatial-temporal forecasting at each edge site.
- The final forecasting results combine both the intra-site and inter-site correlations.



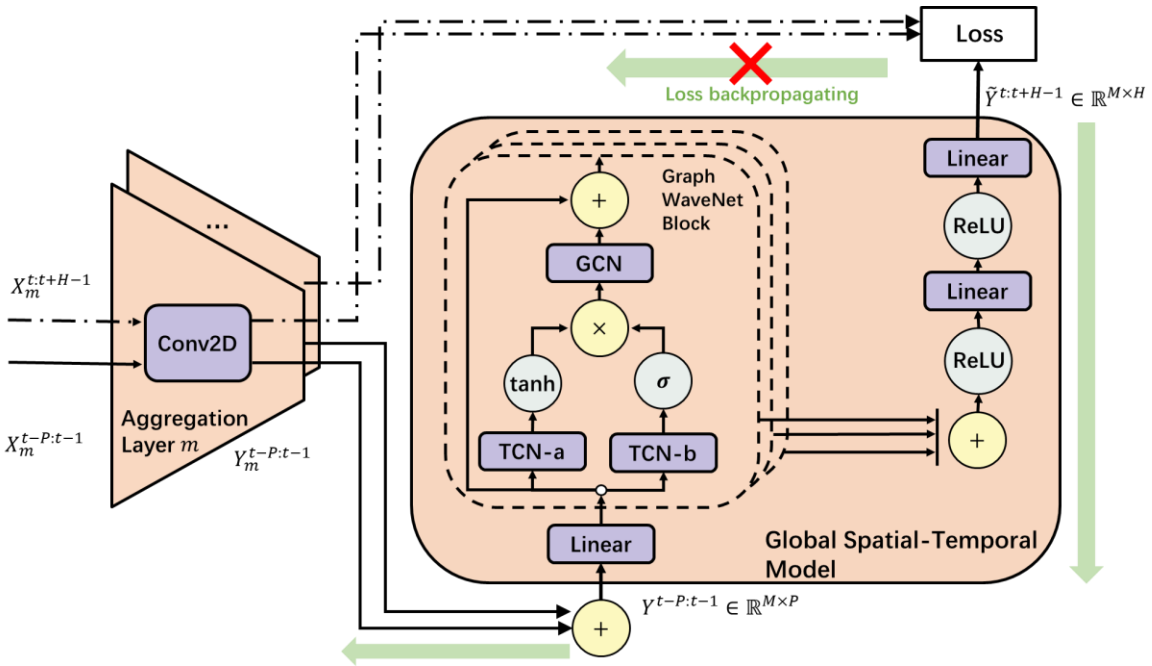
ELASTIC: A Collaborative Cloud-Edge Approach

It mainly consists of two stages:

- The global stage performs coarse-grained spatial-temporal forecasting at the centralized cloud;
- The local stage performs fine-grained spatial-temporal forecasting at each edge site.
- The final forecasting results combine both the intra-site and inter-site correlations.



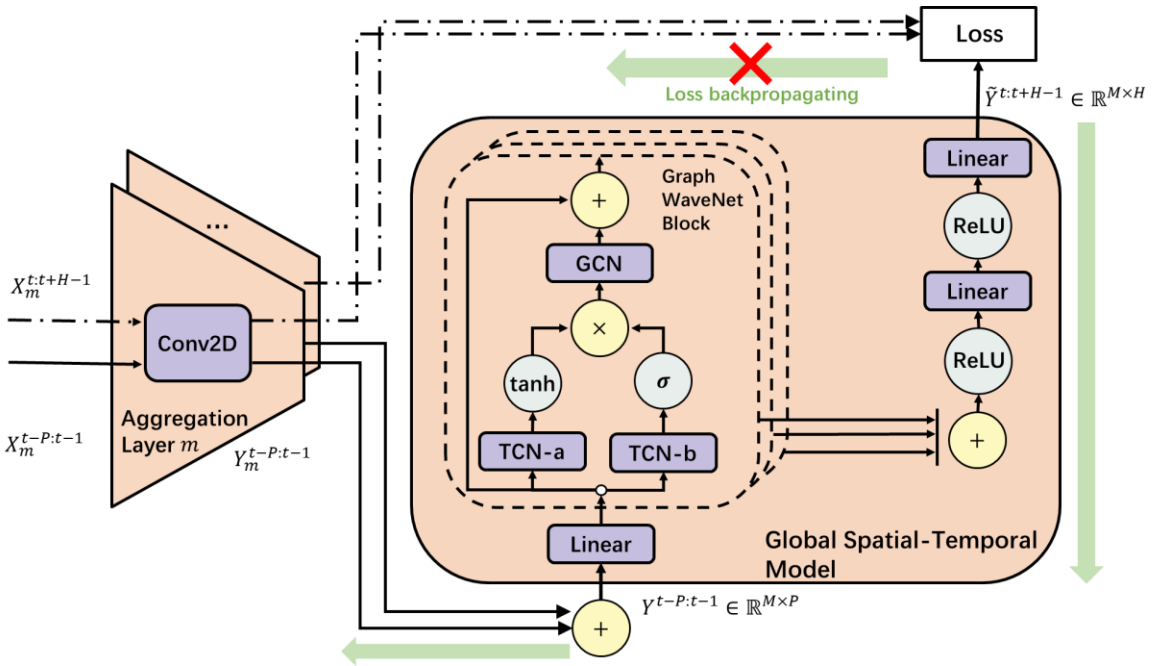
Global Stage



The global stage consists of:

- The aggregation layers on each edge site first aggregate the raw data of each edge site $X_m^{t-P:t-1} \in \mathbb{R}^{N_m \times P}$ into $Y_m^{t-P:t-1} \in \mathbb{R}^{1\times P}$ and then send to the centralized cloud.
- The global spatial-temporal model on the centralized cloud first concatenate the aggregated data together and then capture the inter-site correlations using SOTA methods.

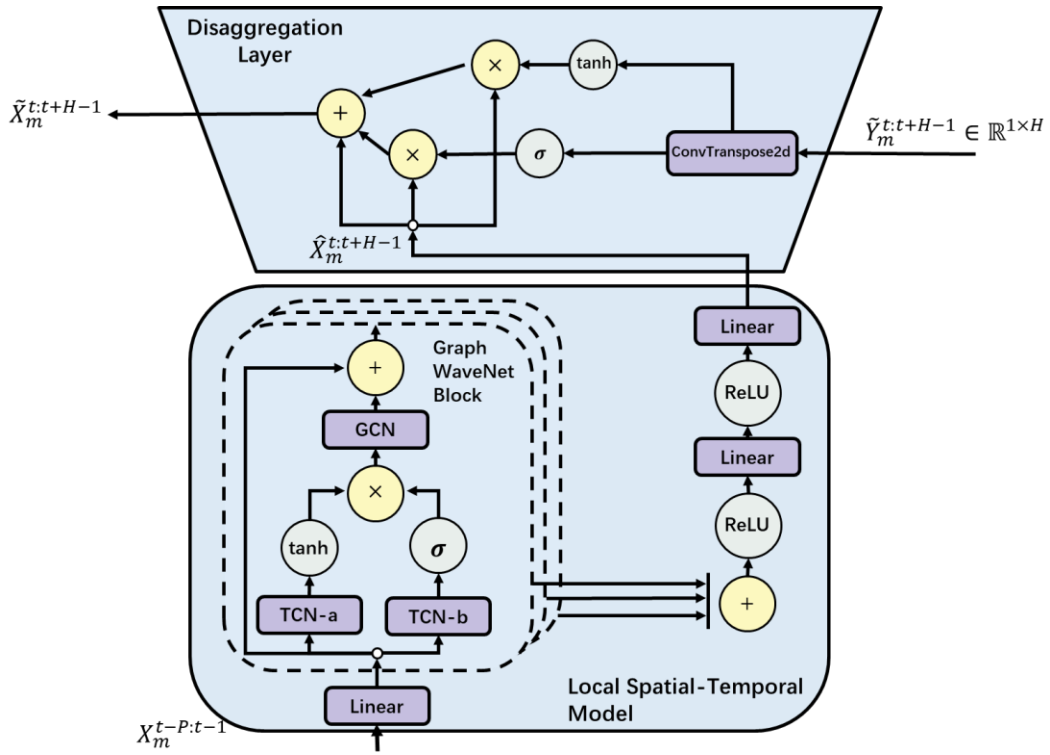
Global Stage



The global stage consists of:

- 1. Decrease time consumption
- 2. Capture intra-site correlation

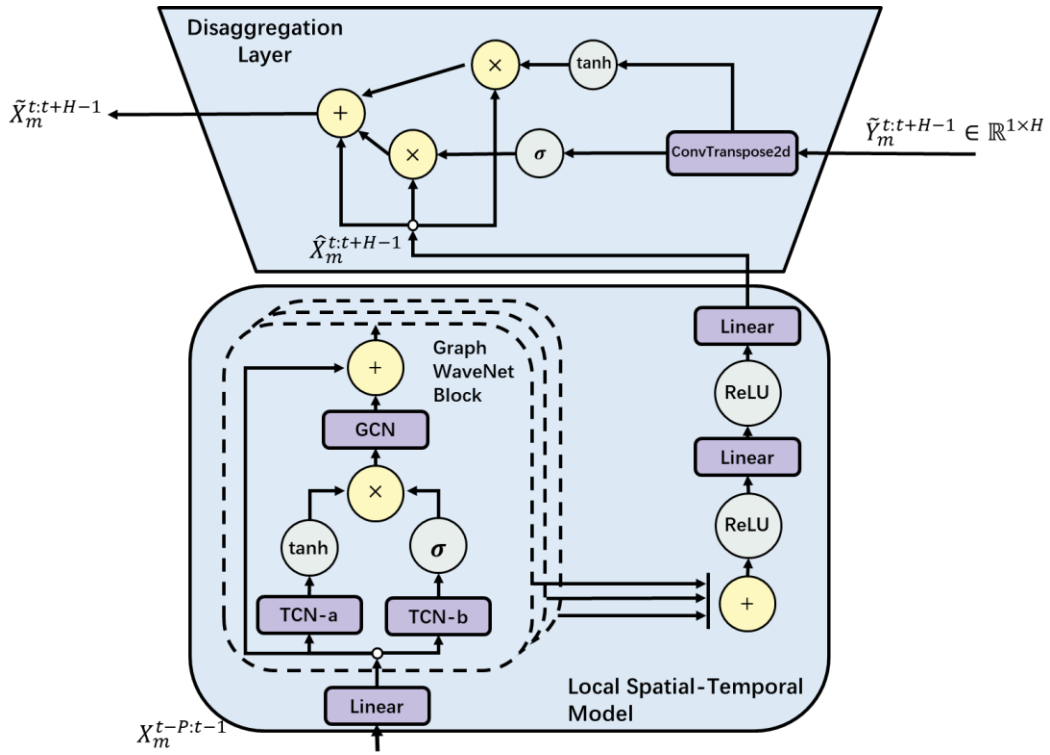
Local Stage



The local stage of each edge site consists of :

- The local spatial-temporal model can utilize SOTA models to capture the intra-site correlations among VMs of same sites.
- The disaggregation layer finally fuse the results of global and local stages in both linear and nonlinear manner.

Local Stage



The local stage of each edge site consists of :

- 1. Capture inter-site correlation
- 2. Increase model accuracy by fusion

Complexity Analysis

- Edge-only methods: $O(\overline{N_m}^2)$;
- Cloud-only methods: $O(N^2)$;
- ELASTIC : $O(N\overline{N_m})$;

ELASTIC is similar to edge-only methods and smaller than cloud-only methods.

where M is the number of edge sites, $N = \overline{N_m} * M$ $(M \ll N)$ is the total number of VMs among M edge sites, and $\overline{N_m}$ is the average number of VMs of each edge site.

Evaluation Setup

Datasets. Realistic workload (CPU and bandwidth) datasets collected from Alibaba ENS, one of the largest ESPs in China [5].

Network environments. To simulate the bandwidth between edge sites and cloud.

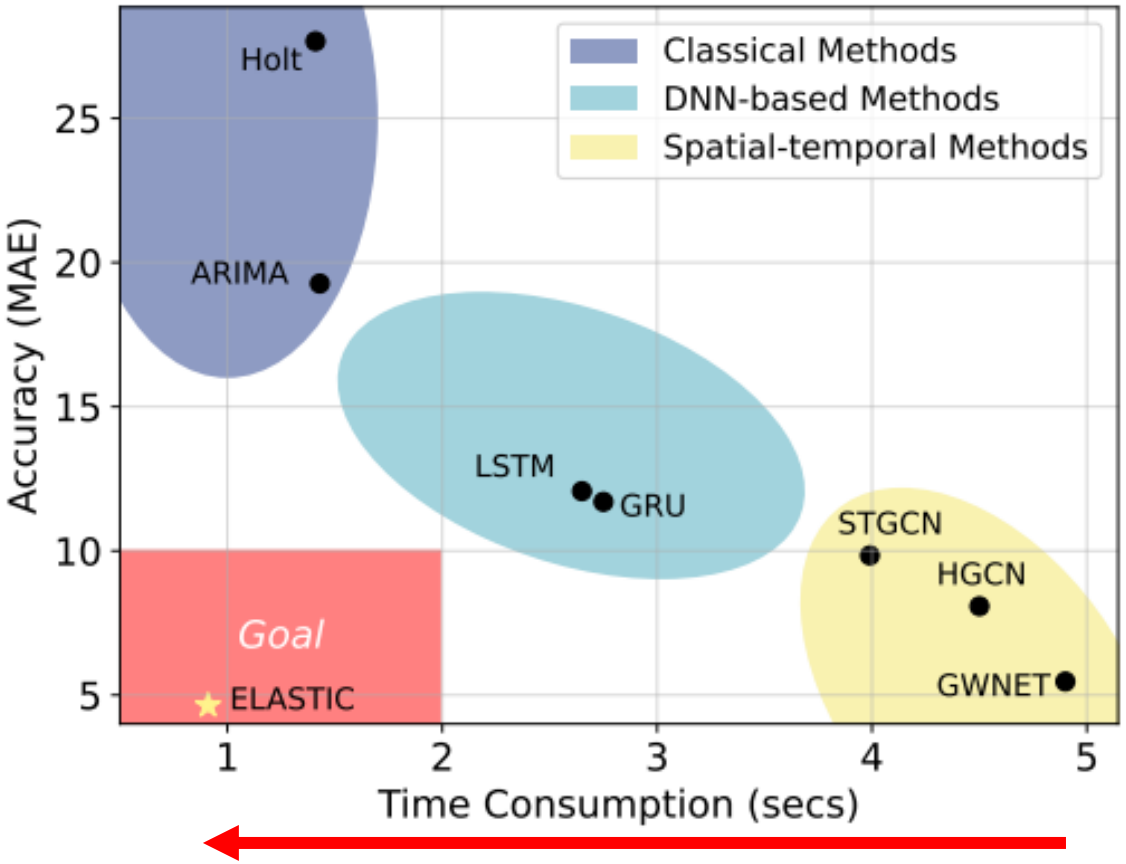
- Slow: 4 Mbps
- Medium: 16Mbps
- Fast: 50 Mbps

Baselines.

- Classical methods: ARIMA, Holt;
- DNN-based methods: LSTM, GRU;
- Spatial-temporal methods: STGCN, HGCN, GWNET.

[5] EdgeWorkloadsTraces. https://github.com/xumengwei/EdgeWorkloadsTraces

Comparison with SOTA Methods

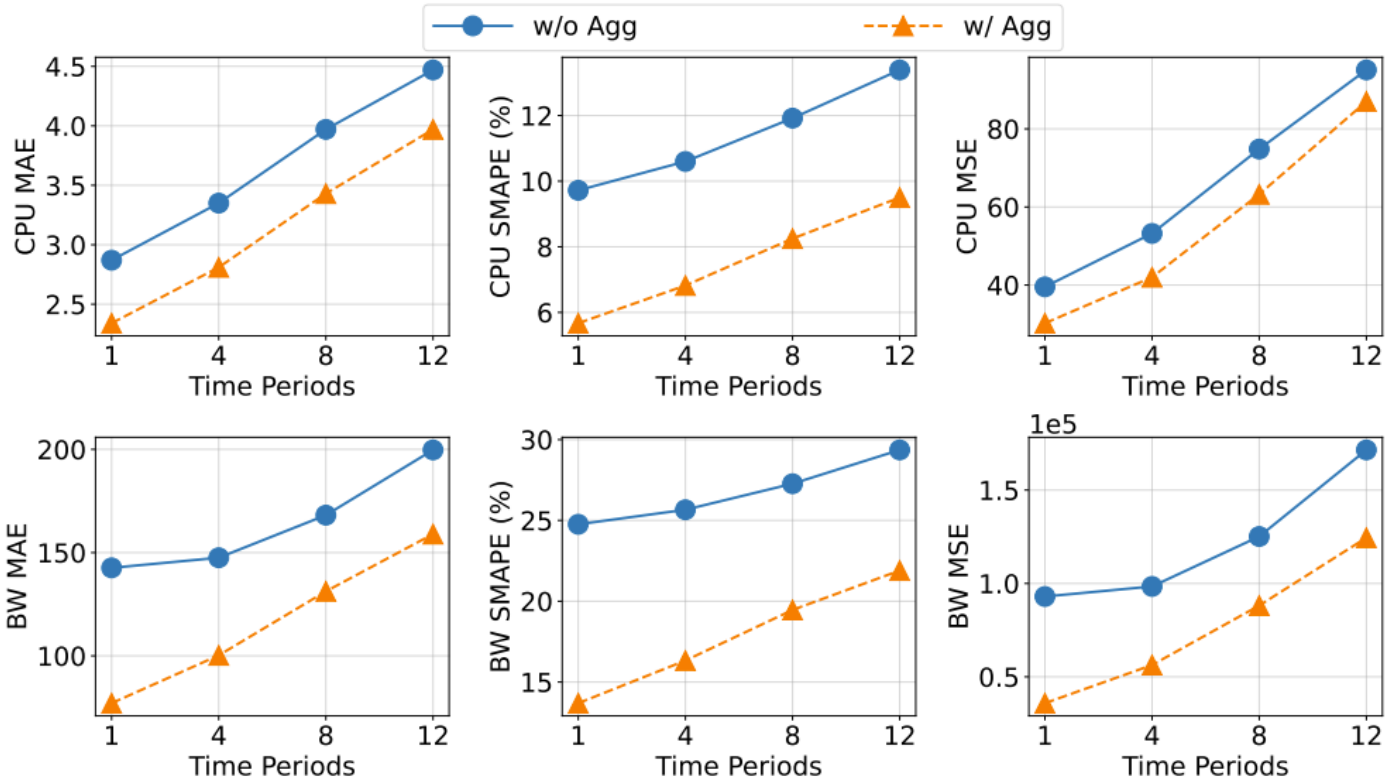


81.19%

81.43%

ELASTIC can achieve better prediction accuracy while the time consumption is significantly reduced.

Ablation Studies of ELASTIC



CPU Datasets:

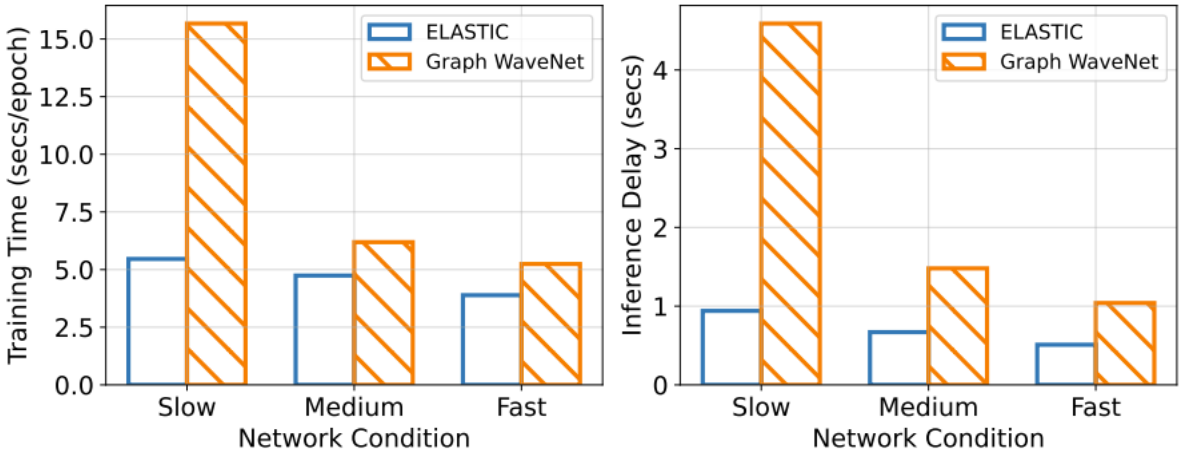
- MAE have decreased 13.98%
- SMAPE have decreased 18.82%
- MSE have decreased 13.60%

BW Datasets:

- MAE have decreased 37.64%
- SMAPE have decreased 10.52%
- MSE have decreased 55.23%

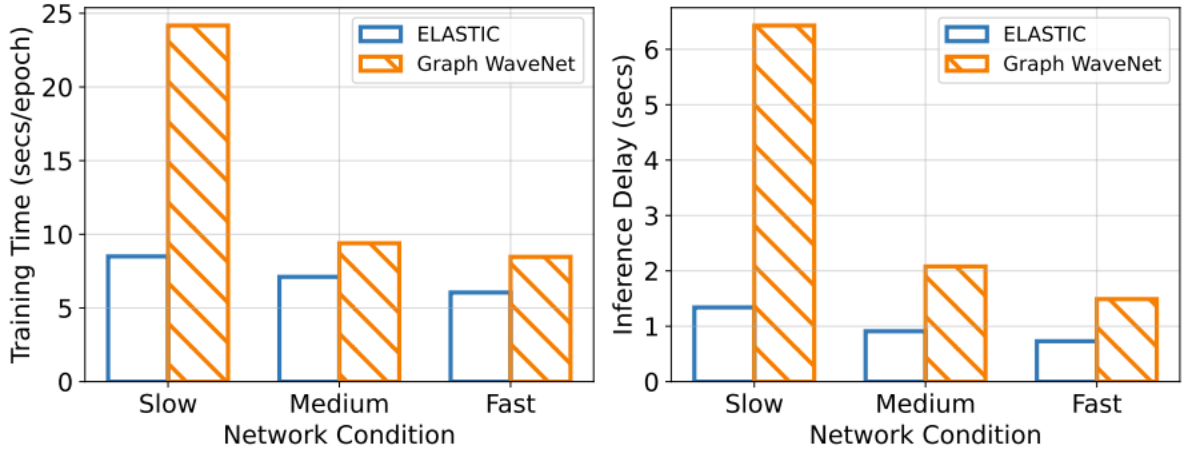
Ablation studies further confirm the effectiveness of different modules.

Time Consumption



CPU Datasets:

- Training time have decreased 38.11%
- Inference time have decreased 61.74%



BW Datasets:

- Training time have decreased 39.23%
- Inference time have decreased 62.14%.

Our Contributions

- This is the first study that leverages the collaborative cloud-edge paradigm for edge workload forecasting.
- We propose ELASTIC, a novel two-stage framework capturing both intra-site and inter-site correlations, which not only increase prediction accuracy but also decrease time consumption.
- Extensive evaluation utilizing the real-world datasets demonstrates the effectiveness of ELASTIC.

Thank You

Yanan Li
YaNanLi@bupt.edu.cn