О методе прогнозирования использования вычислительных ресурсов CPU в облаках

Задорожная Юлия Андреевна[1], **Пашков Василий Николаевич**[2]

[1] Кафедра автоматизации систем и вычислительных комплексов, e-mail: s02190029@gse.cs.msu.ru
[2] Кафедра автоматизации систем и вычислительных комплексов, e-mail: pashkov@lvk.cs.msu.su

Прогнозирование использования CPU является важной и актуальной задачей, так как этот вычислительный ресурс является одним из самых критически важных ресурсов в облачных вычислениях. Если недостаточно ресурсов CPU, приложения могут работать медленно, а пользователи могут столкнуться с задержками и проблемами доступа к сервисам. Если же ресурсов CPU избыток, то есть одна из виртуальных машин сервера работает непрерывно, превышая допустимую нагрузку, а другая имеет минимальную загрузку CPU - это может привести к неоправданным расходам на облачные услуги со стороны провайдера, например, из-за неэффективного использования вычислительного ресурса.

Если ресурсов CPU избыток, это означает, что компьютерная система имеет свободные ресурсы CPU, которые не используются в настоящее время. Это может происходить в различных сценариях, например, если система работает с небольшим количеством задач или если оборудование CPU в системе превышает требования для запущенных задач.

Для облачных провайдеров ключевыми требованиями к прогнозу ресурсов является точность, временное окно прогнозирования, скорость и ресурсоемкость. Знание того, как будет использоваться CPU в будущем, позволяет проводить превентивные меры для балансировки нагрузки, оптимизировать и вовремя масштабировать ресурсы, чтобы обеспечить бесперебойную работу системы и улучшить качество обслуживания клиентов.

Пусть сервер обладает следующим набором физических ресурсов, заданным в виде вектора $H = (V_{CPU}^{host}, V_{RAM}^{host}, V_{SSD}^{host}, \dots)$. Рассмотрим:

— V_{CPU} - объем физических ресурсов CPU сервера;

Пусть на заданном сервере запущено N виртуальных машин (VM). Каждая виртуальная машина характеризуется набором показателей потребления физических ресурсов сервера в каждый момент времени:

$$VM_i\,(t) = \big(v_i^{CPU}\big(\,t), v_i^{RAM}(t), v_i^{SSD}(t), \dots), i = 1..N$$

Таким образом, для поставленной задачи при нормальном режиме функционирования облака, в каждый момент времени суммарное количество потребляемых физических ресурсов сервера всеми запущенными виртуальными машинами на нем не должно превышать его доступных объемов CPU.

Для предотвращения перегрузок и сбоев системы, возникающих при недостатке физических ресурсов, необходимо иметь время для принятия решения о выключении виртуальной машины, ее миграции или для запроса дополнительных ресурсов и подготовкой их к использованию. Это время можно сократить, заранее спрогнозировав потребление ресурсов на некоторый краткосрочный период времени, используя методы глубокого машинного обучения[1].

Задача прогнозирования CPU в облаке в данной работе ведется на основе данных Alibaba Cloud trace 2018 [2], собранных за месяц. Она заключается в использовании исторических данных о загрузке CPU сервера в облачной среде Alibaba Cloud для предсказания будущей нагрузки на CPU в данной среде на краткосрочный период времени. Датасет содержит информацию о ресурсах сервера, использованных в течение дня, включая использование CPU, памяти, сети и диска.

В работе проводится сравнительный анализ существующих подходов и методов краткосрочного прогнозирования использования основных ресурсов сервера виртуальными машинами и их сравнительный анализ, исследуя статистическую модель для анализа временных рядов ARIMA [3], однородные нейронные сети такие, как LSTM [3], GRU [4], а также гибридные нейронные сети в виде LSTM и CNN [5].

В экспериментальной части исследования предлагаемое решение основывается на прогнозирование CPU, используя такие вычислительные ресурсы сервера, как CPU, RAM и загрузка дискового пространства, и сравнение результатов однородных и гибридных методах, состоящих из нейронных сетей (ResNet, 1D-CNN, BI-LSTM) и статистической модели для анализа временных рядов (ETS - Error-Trend-Seasonality). По итогам прогноза формируются стандартные метрики показателя качества прогноза (MAE, RMSE, SMAPE) на несколько шагов дискретного времени вперед.

Анализ, описание и предобработка данных выполнена с помощью Python-библиотек (Pandas, NumPy, Prophet и Statsmodels). Она заключается в исследовании временного ряда CPU, в нахождении зависимостей CPU от других вычислительных ресурсов виртуальной машины. Реализация выбранных методов для экпериментального исследования проведена на языке Python3 с помощью технологий PyTorch, TensorFlow и Keras.

Список литературы

[1] Hsieh S. Y. et al. Utilization-prediction-aware virtual machine consolidation approach for energy-efficient cloud data centers //Journal of Parallel and Distributed Computing. – 2020. – Т. 139. – С. 99-109.

[2] https://github.com/alibaba/clusterdata/tree/master/cluster-trace-v2018

[3] Janardhanan D., Barrett E. CPU workload forecasting of machines in data centers using LSTM recurrent neural networks and ARIMA models //2017 12th international conference for internet technology and secured transactions (ICITST). – IEEE, 2017. – С. 55-60.

[4] Cheng Y. et al. Gru-es: Resource usage prediction of cloud workloads using a novel hybrid method //2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). – IEEE, 2019. – С. 1249-1256.

[5] Ouhame S., Hadi Y., Ullah A. An efficient forecasting approach for resource utilization in cloud data center using CNN-LSTM model //Neural Computing and Applications. – 2021. – Т. 33. – С. 10043-10055.