УДК 004.75:004.8

# ИНТЕЛЛЕКТУАЛЬНОЕ ПРОГНОЗИРОВАНИЕ ЗАГРУЗКИ ОБЛАЧНОЙ ИНФРАСТРУКТУРЫ С ИСПОЛЬЗОВАНИЕМ МОДЕЛИ CHRONOS-2 И МЕТОДОВ АНАЛИЗА ВРЕМЕННЫХ РЯДОВ

***К.А. Афанасенко***
*Магистрант факультета информационных технологий и управления специальности «Системы управления информацией»*
*karnei.afanasenka@gmail.com*

***Н.А. Жиляк***
*Доцент кафедры информационных технологий автоматизированных систем, кандидат технических наук*
*gznadya@gmail.com*

***К.А. Афанасенко***

*Магистрант Белорусского государственного университета информатики и радиоэлектроники. Область научных интересов связана с разработкой алгоритмов поддержки принятия решений, разработкой масштабируемых информационно-компьютерных систем и сервисов.*

***Н.А. Жиляк***

*Доцент кафедры информационных технологий автоматизированных систем, доцент, кандидат технических наук. Область профессиональных и научных интересов связана с информационными системами и технологиями, веб-программированием, веб-проектированием, веб-дизайном, front-end'ом, операционными системами.*

**Аннотация.** В статье рассматривается возможность применения модели Chronos-2 для прогнозирования загрузки облачной инфраструктуры по временным рядам эксплуатационных метрик, включая загрузку центрального процессора, использование оперативной памяти и сетевую активность. Выполнен обзор существующих подходов к прогнозированию временных рядов в облачных системах, включая статистические, нейросетевые и предобученные модели. Показаны особенности модели Chronos-2 как инструмента одномерного, многомерного и covariate-informed прогнозирования. Рассматриваются возможные направления использования такой модели в задачах управления ресурсами, масштабирования сервисов и планирования нагрузки облачной платформы.

**Ключевые слова:** Big Data, облачные вычисления, временные ряды, прогнозирование нагрузки, Chronos-2, оптимизация ресурсов, машинное обучение

**Введение.** Эффективное управление облачной инфраструктурой связано с необходимостью своевременного перераспределения вычислительных ресурсов при изменяющейся нагрузке. Для облачных приложений задача прогнозирования нагрузки рассматривается как важный элемент упреждающего управления ресурсами, поскольку позволяет заранее оценивать будущие изменения спроса на вычислительные мощности, память и сетевые ресурсы [1]. При этом телеметрия облачной платформы формируется в виде временных рядов и включает метрики загрузки CPU, памяти, дисковой подсистемы и сетевого ввода-вывода, что делает задачу прогнозирования непосредственно связанной с методами анализа временных рядов и инструментами Big Data [1].

В последние годы наряду со специализированными моделями для временных рядов получили развитие foundation-модели, обучаемые на больших массивах временных данных. Семейство Chronos основано на языковых архитектурах, в которых временной ряд преобразуется в последовательность токенов, после чего модель обучается по принципу, близкому к обучению языковых моделей [2]. Модель Chronos-2 развивает этот подход и, согласно описанию авторов, поддерживает одномерное, многомерное и covariate-informed прогнозирование в zero-shot режиме, что делает её применимой для задач совместного прогнозирования нескольких взаимосвязанных метрик облачной инфраструктуры [3].

В связи с этим актуальной является задача исследования возможностей модели Chronos-2 для интеллектуального прогнозирования загрузки облачной инфраструктуры. Целью работы является анализ применимости Chronos-2 для предсказания изменения эксплуатационных метрик облачной платформы и рассмотрение возможностей использования такого прогноза в задачах управления ресурсами, автоскейлинга и планирования нагрузки.

**Анализ задачи прогнозирования загрузки облачной инфраструктуры.** Прогнозирование загрузки облачной инфраструктуры рассматривается как одна из ключевых задач упреждающего управления ресурсами. В обзоре по cloud workload prediction отмечается, что точный прогноз нагрузки важен для proactive resource management облачных приложений. На практике телеметрия инфраструктуры формируется в виде временных рядов и включает показатели загрузки процессора, памяти, дисковой подсистемы и сетевого трафика [1], [4].

Особенность этой задачи состоит в том, что нагрузка облачной платформы изменяется неравномерно. Во временных рядах могут одновременно присутствовать периодические колебания, кратковременные всплески, долговременные тренды и нерегулярные изменения, связанные с поведением пользователей, расписанием вычислительных задач и характером развернутых сервисов. Поэтому задача прогнозирования не сводится к простой экстраполяции предыдущих значений и требует применения моделей, способных учитывать сложную структуру зависимостей во времени.

Практическая значимость прогноза определяется тем, что его результаты могут использоваться при масштабировании сервисов, планировании вычислительных мощностей и перераспределении ресурсов между виртуальными машинами и приложениями. При наличии прогноза система управления инфраструктурой может

заранее реагировать на ожидаемый рост или спад нагрузки, а не только фиксировать уже наступившее изменение состояния.

С точки зрения методов анализа данных задача относится к прогнозированию временных рядов. Для нее применяются статистические, машинные и нейросетевые подходы, однако в последние годы появились универсальные предобученные модели для временных рядов. В работе Chronos временной ряд преобразуется в последовательность токенов, после чего используются подходы, близкие к языковому моделированию [5]. В работе Chronos-2 модель описана как zero-shot предобученная модель для временных рядов для одномерного, многомерного и covariate-informed прогнозирования, что представляет интерес для облачной инфраструктуры, где разные метрики часто взаимосвязаны и должны анализироваться совместно.

Таким образом, задача прогнозирования загрузки облачной инфраструктуры является практически значимой задачей анализа временных рядов, связанной с обработкой потоков эксплуатационной телеметрии и поддержкой решений по управлению ресурсами. В рамках данной работы целесообразно рассматривать применение Chronos-2 как современного инструмента прогнозирования, пригодного для анализа нескольких связанных метрик облачной платформы.

**Обзор методов прогнозирования временных рядов.** Для прогнозирования временных рядов традиционно применяются статистические методы, среди которых наиболее распространены экспоненциальное сглаживание и ARIMA. В учебнике Forecasting: Principles and Practice отмечается, что именно эти два подхода относятся к числу наиболее широко используемых в задачах прогнозирования. При этом модели экспоненциального сглаживания описывают тренд и сезонность, а ARIMA ориентирована на моделирование автокорреляционной структуры ряда [6]. Для задач с относительно устойчивой динамикой такие методы остаются полезными, однако при сложных нелинейных зависимостях и большом числе взаимосвязанных признаков их возможности оказываются ограниченными.

С развитием методов глубокого обучения в задачах прогнозирования временных рядов широкое распространение получили рекуррентные нейросетевые архитектуры. Модель LSTM была предложена как способ обучения на последовательностях с длинными временными зависимостями за счет механизма поддержания постоянного потока ошибки во времени [7]. В дальнейшем появились и более специализированные архитектуры. Например, в работе N-BEATS представлена глубокая нейросетевая модель для точечного прогнозирования одномерных временных рядов, ориентированная на универсальность применения и интерпретируемость [8]. Эти методы расширили возможности прогнозирования по сравнению с классическими статистическими моделями, однако в большинстве случаев оставались ориентированными либо на конкретный тип временного ряда, либо на отдельный класс задач.

Следующий этап развития связан с использованием Transformer-подходов. В работе Autoformer предложена архитектура, сочетающая декомпозицию временного ряда на составляющие и механизм Auto-Correlation для выявления зависимостей на длинных интервалах [9]. В модели PatchTST временной ряд разбивается на фрагменты, используемые как входные токены Transformer, что позволяет увеличить длину анализируемого контекста и повысить качество долгосрочного прогнозирования многомерных рядов [10]. Эти модели показали, что механизмы внимания могут быть эффективно адаптированы к временным рядам, особенно в задачах долгосрочного прогноза.

В последние годы внимание сместилось к foundation-моделям для временных рядов. В работе Chronos предложен подход, в котором временной ряд преобразуется в последовательность токенов, после чего используется архитектура, близкая к языковым моделям. Модель Chronos-2 развивает этот подход и описывается авторами как zero-shot

универсальная предобученная модель для временных рядов для одномерного, многомерного и covariate-informed прогнозирования. Для задачи прогнозирования загрузки облачной инфраструктуры это особенно важно, поскольку эксплуатационные метрики, такие как загрузка CPU, использование памяти и сетевой трафик, часто взаимосвязаны и должны анализироваться совместно. Поэтому в рамках данной работы именно модели класса Chronos представляют наибольший интерес как современные универсальные средства прогнозирования временных рядов .

**Применение модели Chronos-2 для прогнозирования нагрузки облачной инфраструктуры.** Модель Chronos была предложена как foundation-подход к прогнозированию временных рядов, в котором значения ряда предварительно масштабируются, квантуются и преобразуются в последовательность токенов, после чего для обучения используются архитектуры, основанные на Transformer-моделях языкового типа. Такой подход позволяет рассматривать прогнозирование временных рядов как задачу, близкую к последовательностному моделированию, и использовать преимущества предобученных моделей при работе с новыми рядами без построения отдельной специализированной архитектуры под каждую прикладную задачу [2].

В работе Chronos-2 этот подход был расширен от одномерного прогнозирования к более универсальному сценарию. Авторы описывают Chronos-2 как предобученную zero-shot модель, поддерживающую одномерное, многомерное и covariate-informed прогнозирование. Это особенно важно для задач облачной инфраструктуры, поскольку эксплуатационная нагрузка обычно не сводится к одному показателю: загрузка CPU, использование памяти, сетевой трафик и другие метрики изменяются совместно и могут оказывать влияние друг на друга. Поэтому для данной предметной области более уместно рассматривать не изолированный прогноз одной метрики, а совместный анализ нескольких связанных временных рядов.

Chronos-2 использует механизм group attention, который поддерживает обмен информацией между временными рядами внутри одной группы, включая связанные ряды, разные переменные многомерного ряда, а также целевые признаки и ковариаты. Для облачной платформы это означает возможность учитывать не только историю самой прогнозируемой метрики, но и дополнительные признаки, связанные с поведением системы. В прикладной постановке это может соответствовать совместному использованию рядов CPU, RAM, Network I/O и, при необходимости, внешних факторов, если они доступны в данных мониторинга.

Еще одной важной особенностью Chronos-2 является ориентация на zero-shot использование, то есть получение прогноза без обязательного дообучения под каждую новую задачу. Для облачной аналитики это представляет практический интерес, поскольку позволяет применять модель в условиях, где инфраструктурные конфигурации, профили сервисов и состав наблюдаемых метрик могут меняться. В таком случае модель может рассматриваться как универсальный инструмент предварительного прогноза, который затем используется в контуре принятия решений по масштабированию и распределению ресурсов.

**Подход к формированию входных данных и прогнозируемых метрик.** В задаче прогнозирования загрузки облачной инфраструктуры входные данные целесообразно формировать на основе эксплуатационной телеметрии, поступающей от виртуальных машин, контейнеров и сервисов мониторинга. К числу основных метрик обычно относят загрузку центрального процессора, использование оперативной памяти, показатели дисковой подсистемы и сетевой трафик. Именно такие данные рассматриваются как базовые показатели производительности облачной среды и используются при анализе состояния инфраструктуры.

В рамках предлагаемого подхода в качестве прогнозируемых величин можно рассматривать как отдельную метрику, например загрузку CPU, так и совокупность

взаимосвязанных показателей. Второй вариант представляется более содержательным для облачной инфраструктуры, поскольку изменение нагрузки на процессор, память и сеть в реальных системах часто происходит не изолированно, а совместно. В этом случае модель получает не один временной ряд, а набор синхронизированных рядов, описывающих состояние платформы на каждом временном шаге. Это соответствует заявленным возможностям Chronos-2 для многомерного и covariate-informed прогнозирования.

Для использования таких данных необходимо привести их к единой временной сетке. Это означает, что значения всех метрик должны быть согласованы по шагу наблюдения, например по минутам, пяти минутам или часам. После этого данные могут быть агрегированы, очищены от пропусков и явных выбросов, а затем представлены в виде последовательностей фиксированной длины.

Входной фрагмент должен содержать историю изменения метрик за некоторый предшествующий интервал, а выходом модели должен становиться прогноз на ближайший горизонт времени. Такой способ представления позволяет использовать временной контекст и подготавливает данные к подаче в модель временных рядов.

**Практическое применение прогнозирования в задачах управления облачной инфраструктурой.** Прогнозирование нагрузки представляет практический интерес для систем управления облачной инфраструктурой, поскольку телеметрические данные используются при оценке производительности и планировании ресурсов, а в обзоре по cloud workload prediction сама задача рассматривается как часть proactive resource management. Для облачной платформы это означает, что прогнозируемые значения эксплуатационных метрик могут применяться не только для анализа текущего состояния, но и для подготовки инфраструктуры к ожидаемому изменению нагрузки.

В прикладной постановке прогноз может использоваться в нескольких направлениях. Во-первых, он может служить основой для предварительного масштабирования вычислительных ресурсов при ожидаемом росте нагрузки. Во-вторых, он может использоваться при перераспределении ресурсов между виртуальными машинами, контейнерами или сервисами, если по ряду метрик наблюдается устойчивое приближение к перегрузке. В-третьих, прогноз может быть полезен при планировании снижения резервирования в периоды ожидаемого спада активности.

Такая логика соответствует общему подходу к упреждающему управлению ресурсами, описываемому в работах по прогнозированию нагрузки облачных систем.

Для данной задачи особый интерес представляет возможность совместного анализа нескольких метрик. Chronos-2 описывается как zero-shot модель для одномерного, многомерного и covariate-informed прогнозирования, а также использует механизм group attention для обмена информацией между связанными временными рядами внутри группы. Поэтому в облачной инфраструктуре модель может рассматриваться как средство совместного прогноза CPU, памяти, сетевой активности и других взаимосвязанных показателей, а не только одного изолированного ряда. Это делает подход более содержательным для задач управления платформой, где решение обычно принимается не по одной метрике, а по совокупности признаков состояния системы.

Дополнительное прикладное значение связано с zero-shot характером модели Chronos-2. Авторы позиционируют ее как универсальную предобученную модель, пригодную для использования без обязательного дообучения под каждую новую задачу прогнозирования [3]. Для облачной среды это может быть полезно в случаях, когда инфраструктурная конфигурация изменяется, появляются новые сервисы или набор доступных метрик отличается между площадками.

В такой постановке модель может использоваться как общий инструмент прогностической аналитики, который встраивается в контур мониторинга и поддержки решений по управлению ресурсами.

**Заключение.** В работе рассмотрена задача интеллектуального прогнозирования загрузки облачной инфраструктуры на основе анализа временных рядов эксплуатационных метрик.

Рассмотрено, что телеметрия облачной платформы формируется в виде последовательностей значений загрузки процессора, памяти, дисковой подсистемы и сетевой активности, а потому задача управления ресурсами тесно связана с методами прогнозирования временных рядов.

Проведённый обзор показывает, что наряду с традиционными статистическими и нейросетевыми подходами существенный интерес представляют предобученные модели для временных рядов.

Модели семейства Chronos основаны на представлении временного ряда в виде последовательности токенов, а Chronos-2 расширяет этот подход до одномерного, многомерного и covariate-informed прогнозирования в zero-shot постановке.

Для задач облачной инфраструктуры это особенно важно, поскольку эксплуатационные метрики обычно взаимосвязаны и должны рассматриваться совместно.

На основании рассмотренных материалов можно отметить, что модель Chronos-2 представляет методический интерес для задач упреждающего управления облачными ресурсами.

Такая модель может рассматриваться как перспективная основа для прогнозирования изменения нагрузки и последующего применения прогноза в задачах масштабирования, перераспределения ресурсов и планирования эксплуатации облачной платформы.

Перспективы дальнейшей работы связаны с уточнением состава входных метрик, формированием согласованных многомерных рядов облачной телеметрии и последующей оценкой применимости модели Chronos-2 в конкретных сценариях управления инфраструктурой.

**Список литературы**

[1] Feng B., et al. Application-Oriented Cloud Workload Prediction: A Survey. 2025. DOI: 10.26599/TST.2024.9010024.

[2] Ansari A.F., et al. Chronos: Learning the Language of Time Series. Amazon Science. 2024. DOI: 10.48550/arXiv.2510.15821.

[3] Ansari A.F., et al. Chronos-2: From Univariate to Universal Forecasting. arXiv. 2025. DOI: 10.48550/arXiv.2510.15821.

[4] Architecture strategies for collecting performance data. Microsoft Learn.URL: https://learn.microsoft.com/en-us/azure/well-architected/performance-efficiency/collect-performance-data.

[5] Adapting language model architectures for time series forecasting. Amazon Science Blog. URL: https://www.amazon.science/blog/adapting-language-model-architectures-for-time-series-forecasting.

[6] Hyndman R.J., Athanasopoulos G. Forecasting: Principles and Practice. 3rd ed. Melbourne: OTexts; 2018.

[7] Hochreiter S., Schmidhuber J. Long Short-Term Memory. Neural Computation. 1997;9(8):1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

[8] Oreshkin B.N., Carpov D., Chapados N., Bengio Y. N-BEATS: Neural Basis Expansion Analysis for Interpretable Time Series Forecasting. ICLR. 2020. DOI: 10.48550/arXiv.1905.10437.

[9] Wu H., Xu J., Wang J., Long M. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. NeurIPS. 2021. DOI: 10.48550/arXiv.2106.13008.

[10] Nie Y., Nguyen N.H., Sinthong P., Kalagnanam J. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. ICLR. 2023. DOI: https://doi.org/10.48550/arXiv.2211.14730.

**Авторский вклад**

**Афанасенко Корней Александрович** – постановка задачи исследования, руководство исследованием, анализ существующих решений, изучение возможных проектных решений.

**Жиляк Надежда Александровна –** научное руководство исследованием, формирование направления работы и постановки задач, критический анализ содержания работы и редактирование научного текста

# INTELLIGENT FORECASTING OF CLOUD INFRASTRUCTURE LOAD USING THE CHRONOS-2 MODEL AND TIME SERIES ANALYSIS METHODS

***K.A. Afanasenko***
*Master's student in the Faculty of Information Technologies and Control, majoring in "Information Management Systems"*

***N.A. Zhilyak***
*Associate Professor, Department of Information Technologies for Automated Systems, Candidate of Technical Sciences*

**Abstract.** This article discusses the application of the Chronos-2 model for predicting the dynamics of computing resource utilization, including CPU load, RAM usage, and network activity. The use of the Chronos-2 model is proposed for predicting the dynamics of computing resource utilization, including CPU load, RAM usage, and network activity. An analysis of existing approaches to time series forecasting in cloud systems, including statistical and neural network methods, is conducted. It is described that the use of the Chronos-2 model allows for improved forecasting accuracy compared to classical models and can also be used to optimize resource allocation and enhance the operational efficiency of cloud platforms.

**Keywords**: Big Data, cloud computing, time series, load forecasting, Chronos-2, resource optimization, machine learning.