**УДК** 004.75

# Сравнительный анализ методов предиктивного масштабирования микросервисов с учетом требований SLA и стоимости вычислений

**Давыдов Дмитрий Андреевич**

*ФГБОУ ВО «Казанский национальный исследовательский технический университет им. А.Н. Туполева – КАИ», г. Казань*

После перехода корпоративных информационных систем на микросервисную архитектуру встала задача динамического управления вычислительными ресурсами. Реактивные механизмы оркестраторов, например стандартный Horizontal Pod Autoscaler (HPA), опираются на пороговые значения утилизации CPU и памяти. Такой подход инерционен: при burst-нагрузках система не успевает выделить дополнительные мощности, и метрики фиксируют деградацию времени отклика и нарушение соглашений об уровне обслуживания (SLA). Предиктивные алгоритмы должны устранять этот недостаток, но часто приводят к избыточному резервированию ресурсов (overprovisioning). При оплате по факту потребления это означает неоправданные расходы.

В работе классифицированы и сравнены алгоритмы предиктивного автоскейлинга по балансу между соблюдением SLA и минимизацией стоимости инфраструктуры, а также проанализированы существующие подходы.

Первый класс методов строится на классических моделях временных рядов (ARIMA, экспоненциальное сглаживание). Они вычислительно просты, поэтому их можно встроить прямо в цикл принятия решений контроллера масштабирования. Но при резких немонотонных всплесках трафика, вызванных внешними факторами (например, днями выплаты зарплаты в банковских системах или акциями в интернет-магазинах), ошибка прогнозирования (MAPE) у статистических моделей высокая.

Второй класс образуют подходы на основе глубокого обучения, в частности рекуррентных нейронных сетей (LSTM, GRU) и трансформерных архитектур. Нейросетевые модели выявляют скрытые нелинейные зависимости в потоках телеметрии. Согласно [1], использование LSTM снижает процент

нарушений SLA до 99% при циклических нагрузках. Однако инференс модели сам по себе дорог и требует отдельных узлов кластера, что частично сводит на нет экономический эффект от оптимизации распределения ресурсов [2].

Третий класс составляют гибридные архитектуры на основе обучения с подкреплением (Reinforcement Learning). Агент в среде Kubernetes учится напрямую минимизировать целевую функцию, которая объединяет штраф за нарушение SLA и стоимость инстансов [3]. Методы на базе RL лучше всех адаптируются к нестационарным средам, но требуют долгого обучения и сложной настройки гиперпараметров.

Алгоритмы оценивались по трем метрикам: точность прогноза, процент нарушений SLA и относительная стоимость владения (TCO). Наиболее перспективными оказались многокритериальные модели с экзогенными переменными. Если добавить в модели прогнозирования контекстные данные (календарь бизнес-событий, спотовые цены на облачные ресурсы), MAPE в периоды аномальных нагрузок снижается на величину от 18 до 22% по сравнению с базовыми реализациями KEDA [4].

Результаты можно использовать при проектировании подсистем автоматического масштабирования в цифровых экосистемах. Дальнейшие исследования целесообразно направить на легковесные архитектуры предиктивных контроллеров, которые учитывают бизнес-контекст и не вносят критических задержек в контур управления кластером.

**Список использованной литературы**

1. Долматов Роман Александрович, Сараджишвили Сергей Эрикович ОПТИМИЗАЦИЯ МАСШТАБИРУЕМОСТИ И ОТКАЗОУСТОЙЧИВОСТИ МИКРОСЕРВИСОВ С ПРИМЕНЕНИЕМ KUBERNETES // SAEC. 2024. №2. URL: https://cyberleninka.ru/article/n/optimizatsiya-masshtabiruemosti-i-otkazoustoychivosti-mikroservisov-s-primeneniem-kubernetes (дата обращения: 27.09.2026).

2. В. М. Тхай Энергоэффективное автомасштабирование облачных сервисов с гарантиями хвостовой задержки SLO: подход на основе

ограниченного обучения с подкреплением и защитного слоя // ИВД. 2026. №5 (137). URL: https://cyberleninka.ru/article/n/energoeffektivnoe-avtomasshtabirovanie-oblachnyh-servisov-s-garantiyami-hvostovoy-zaderzhki-slo-podhod-na-osnove-ogranichennogo (дата обращения: 28.09.2026).

3. Н. Б. Лазарева Автоматическое горизонтальное масштабирование приложений в среде Кубернетес // ИВД. 2026. №4 (136). URL: https://cyberleninka.ru/article/n/avtomaticheskoe-gorizontalnoe-masshtabirovanie-prilozheniy-v-srede-kubernetes (дата обращения: 29.09.2026).

4. Рамазанов Н. М. Механизмы поддержания высокой производительности образовательных платформ в условиях пиковых сезонных нагрузок // Системный анализ в науке и образовании. 2025. №1. URL: https://cyberleninka.ru/article/n/mehanizmy-podderzhaniya-vysokoy-proizvoditelnosti-obrazovatelnyh-platform-v-usloviyah-pikovyh-sezonnyh-nagruzok (дата обращения: 29.09.2026).