DOI 10.37882/2223-2966.2026.01.22

ГИБРИДНЫЕ МЕТОДЫ ПРОГНОЗИРОВАНИЯ ЭВОЛЮЦИИ КОМПОНЕНТНОЙ БАЗЫ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ НА ОСНОВЕ ИНТЕГРАЦИИ СТАТИСТИЧЕСКИХ И НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ

HYBRID METHODS FOR PREDICTING THE EVOLUTION OF THE COMPONENT BASE OF COMPUTER SYSTEMS BASED ON THE INTEGRATION OF STATISTICAL AND NEURAL NETWORK MODELS

N. Maximenko

Summary. Transition to heterogeneous nodes, chiplet integration, and coherent buses (CXL, PCIe 6.0) increases the cost of predictive errors in Perf/W, bandwidth, and reliability, while linear methods (ARIMA) are limited in nonlinear dynamics.

The goal is to create a hybrid forecasting methodology for 1–5 year horizons through integration of SARIMA+LSTM+Transformer+GNN with Bayesian calibration and explainability (SHAP).

Methods include robust-scaling, exogenous features (process technology, interfaces), quantile modeling, and rolling-origin validation.

Results.The hybrid ensemble reduces RMSE by 15–24 % and sMAPE by 20–30 % relative to baseline SARIMA; intervals are calibrated (PICP=0.90–0.95, PINAW=0.15–0.20). Superiority is statistically significant (DM, p<0.05).

Keywords: forecasting, hybrid models, ARIMA/SARIMA, LSTM, Transformer for Time Series, graph neural networks, explainable AI, system analysis, Perf/W, SLA.

Максименко Наталья Сергеевна
старший преподаватель, ФГБОУ ВО Донецкий национальный технический университет
Nataly.Maksime@yandex.ru

Аннотация. Переход к гетерогенным узлам, чиплетной интеграции и когерентным шинам (CXL, PCIe 6.0) повышает цену прогностических ошибок по Perf/W, пропускной способности, латентности и надёжности, а линейные методы (ARIMA) ограничены при нелинейной динамике.

Цель — создать гибридную методологию прогнозирования на горизонтах 1–5 лет через интеграцию SARIMA+ARIMAX+LSTM+Transformer for TS+GNN с байесовской калибровкой и explainability (SHAP/LIME).

Методы включают robust-масштабирование, экзогенные признаки (техпроцесс, интерфейсы), квантильное моделирование и rolling-origin валидацию.

Результаты. Гибридный ансамбль снижает RMSE на 15–24 % и sMAPE на 20–30 % относительно базовой SARIMA; интервалы калиброваны (PICP=0,90–0,95, PINAW=0,15–0,20). Превосходство статистически значимо (DM, p <0,05).

Ключевые слова: прогнозирование, гибридные модели, ARIMA/SARIMA, LSTM, Transformer for Time Series, графовые нейросети, explainable AI, системный анализ, Perf/W, SLA.

Введение

Современная вычислительная инженерия переживает трансформацию: исчерпание классических резервов масштабирования сочетается с быстрым ростом потребностей ИИ и высокопроизводительных вычислений. Ответом стали гетерогенные узлы (CPU–GPU–NPU), чиплетная интеграция и когерентные шины (CXL, PCIe 6.0, UCIe), которые радикально меняют траектории эволюции компонентной базы и повышают цену прогностических ошибок по Perf/W, пропускной способности, латентности и надёжности [1]–[4].

Классические методы (ETS, ARIMA) остаются стандартом благодаря прозрачности и низкой вычислительной стоимости [5]–[7], но архитектурные переходы, нелинейные режимы и топологическая сцепленность подсистем (CPU↔Memory↔Interconnect) выходят за пределы их выразительности [8], [9]. Нейросетевые подходы (LSTM, Transformer, GNN) демонстрируют высокую точность на нелинейных процессах [10], [11], однако их внедрение требует обеспечения воспроизводимости, устойчивости и объяснимости [12].

Исследовательский разрыв состоит в отсутствии единого подхода, который одновременно учитывает мультишкальную динамику и топологию, сохраняет интерпретируемость и устойчив к неполноте данных. Настоящая работа закрывает этот разрыв посредством гибридной методологии, объединяющей статистические и нейросетевые модели с байесовской интеграцией экспертной информации.

Цель и задачи исследования

Постановка проблемы. Гетерогенизация платформ и внедрение новых интерфейсов изменяют динамику ключевых метрик: Perf/W, пропускной способности, латентности, надёжности. Для планирования необходимы воспроизводимые прогнозы на горизонтах 1–5 лет, устойчивые к неполноте данных и структурным сдвигам.

Цель исследования. Цель исследования. Разработать и верифицировать гибридную методологию прогнозирования метрик компонентной базы, которая интегрирует ARIMA/SARIMAX, глубокие архитектуры (LSTM, Transformer), GNN и байесовскую калибровку; обеспечивает интерпретируемость (SHAP/LIME) и калиброванные интервалы; демонстрирует статистически значимое превосходство по RMSE/sMAPE.

Формализация цели. Прогнозирование целевого вектора метрик $y_{t+h} \in R^m$ на горизонте $h \in \{1,3,5\}$ лет формулируется как задача аппроксимации отображения

$$\hat{y}_{t+h} = F_\theta(z_t, H_t), \quad (1)$$

где z_t — вектор признаков (техпроцесс, архитектура, тип/поколение памяти, интерфейсы, энергетические режимы), H_t — информационный набор (история, производные признаки и сценарные индикаторы), θ — параметры/гиперпараметры гибридной модели. Отображение (1) задаёт требуемую функциональность: консистентный прогноз по гетерогенным источникам с поддержкой сценариев.

Оптимизация параметров гибридной модели на временных разбиениях (rolling-origin, out-of-time) рассматривается как минимизация взвешенной ожидаемой потери

$$\theta = \arg\min_\theta E\left[\sum_{j=1}^{m} w_j L_j\left(y_{t+h}^{(j)}, \hat{y}_{t+h}^{(j)}\right)\right], \sum_j w_j = 1, \quad (2)$$

где $L_j \in \{RMSE, MAE, sMAPE, pinball\}$ — функции потерь по метрикам j, w_j — веса, отражающие приоритеты KPI (Perf/W, латентность, энергоёмкость). Формулировка (2) фиксирует *количественную трактовку цели*.

Обозначения: t — время (квартал/год), h — горизонт прогноза, m — число целевых метрик, $E[\cdot]$ — усреднение по временным окнам/сценариям, $\hat{y}$ — прогноз.

Исследовательские вопросы и гипотезы:

— H1: гибридный ансамбль снижает RMSE ≥15 % и sMAPE ≥10 % относительно SARIMA на горизонтах 1–5 лет (p <0,05).

— H2: кодирование архитектурных переходов как экзогенных признаков уменьшает ошибки и повышает переносимость между подсистемами.

— H3: байесовская интеграция экспертных квантилей обеспечивает PICP≥0,90 при снижении PINAW.

Задачи исследования. Сформировать корпус данных (2010–2025) по CPU/GPU/памяти/накопителям/межсоединениям с нормализацией и очисткой; спроектировать конвейер предобработки; разработать гибридный ансамбль; интегрировать экспертные квантили; организовать протокол валидации; обеспечить explainability; подготовить визуально-аналитическую отчётность; продемонстрировать практическую полезность.

Для сопоставимости компонентов цели и шагов реализации рассмотрим это в табл. 1. Наконец, чтобы связать формальные метрики качества с прикладными KPI эксплуатации, рассмотрим табл. 2.

Таблица 1.

Карта соответствия цели, задач, методов, артефактов и критериев приёмки

Компонент цели	Ключевые задачи (шаги)	Методы/модели	Артефакты и критерии приёмки
Точность прогнозов	Конвейер данных; ансамбль; настройка гиперпараметров	ETS/ARIMA/ARIMAX; GBM; LSTM/Transformer; GNN	RMSE/sMAPE/MASE; DM ($p < 0,05$); выигрыш ≥ 15/10 %
Калиброванность интервалов	Квантильные модели; бутстрэп остатков	pinball-loss; CRPS; бутстрэп	PICP 80/90/95 % в пределах $\pm 0,02$; минимальный PINAW
Explainability	SHAP/LIME; сценарные What-If	SHAP/LIME; декомпозиции ETS/ARIMA	Карты важности; стабильность рангов на OOT
Воспроизводимость	rolling-origin/OOT; контроль дрейфа; версии данных/моделей	PSI/MMD; протокол разбиений	Репликация; триггеры переподготовки; журналы экспериментов

Для наглядности табл. 1 группирует целевые аспекты по четырём осям и привязывает их к конкретным артефактам приёмки. Такой формат облегчает трассируемость: каждому блоку соответствует измеримый набор метрик и процедур, что снижает риск расхождений между ожиданиями и фактическими результатами.

Чтобы связать формальные метрики качества с инженерными решениями, для наглядности рассмотрим табл. 2.

Таблица 2.
Привязка метрик качества модели к прикладным KPI

Метрика модели	Интерпретация и чувствительность	KPI и инженерная трактовка
RMSE / MAE	Абсолютные ошибки; RMSE усиливает крупные промахи	Планирование Perf/W и энергомощности; резервы охлаждения
sMAPE / MASE	Нормированные относительные ошибки; сопоставимость рядов	Сравнение подсистем (CPU/GPU/Memory/IO); SLA-отчётность
PICP / PINAW	Покрытие и ширина предиктивных интервалов	Риск нарушения SLA-латентности; ширина безопасных коридоров
CRPS / LogScore	Интегральное качество распределений; хвостовые риски	Принятие решений под неопределённостью; устойчивость к шокам

Как видно из табл. 2, каждая группа метрик дополняет другую: абсолютные ошибки важны для бюджетов мощности, нормированные — для сопоставимости подсистем, интервальные — для управления рисками, а распределительные — для сценариев с асимметричными штрафами. В совокупности это формирует прозрачную связь между статистическим качеством и эксплуатационными KPI, что критично при масштабировании решения.

Методы и подходы

Методологическая рамка. Подход объединяет статистические модели временных рядов, методы машинного/глубокого обучения и байесовскую интеграцию экспертов: ETS/Holt–Winters и ARIMA/SARIMAX для уровней, тренда, сезонности; градиентный бустинг (GBM), LSTM/GRU, Transformer для нелинейностей; GNN на графе подсистем; байесовская интеграция экспертных квантилей; причинная валидация: rolling-origin и out-of-time, бутстрэп остатков, калибровка интервалов.

Таксономия методов. Для компактного сопоставления ролей и рисков см. табл. 3.

Таблица 3.
Таксономия методов и их роль в исследовании

Класс метода	Целевая функция/ задача	Сильные стороны	Риски/ ограничения
ETS/ARIMA/ ARIMAX	Краткосрочные горизонты; сезонность; экзогенные факторы	Прозрачность, воспроизводимость, низкая стоимость	Деградация при изломах режима, слабая нелинейность
GBM (градиентный бустинг)	Нелинейные взаимодействия признаков и сценариев	Устойчивость к шуму; SHAP-объяснимость	Переобучение; тюнинг гиперпараметров
LSTM/ Transformer (TS)	Дальние контексты; мультишкальные паттерны; квантили	Высокая точность на сложной динамике	Требовательность к данным; контроль утечек по времени
GNN (графовые НС)	Топология CPU–Memory–IO; каскадные эффекты	Учёт структурных связей и узких мест	Сложность построения/верификации графа
Байесовская интеграция	Калибровка интервалов; устойчивость к малым выборкам	Полнокровные покрытия; учёт априори	Чувствительность к весам и формам априори

После табл. 3 сразу видно комплементарность: статистика задаёт проверяемые «опоры», DL/GNN закрывают нелинейность и топологию, а байесовский слой стабилизирует интервальные характеристики на малых выборках.

Предобработка и причинность. Конвейер предобработки обеспечивает сопоставимость рядов и исключает утечки по времени. — Выравнивание календарей, синхронизация частот, дедупликация источников. — Робастное масштабирование (медиана/MAD), стабилизация дисперсии (Box–Cox). — Диагностика стационарности (ADF/KPSS), сезонности (ACF/PACF), изломов (Chow/Bai–Perron). — Строгая причинность: генерация скейлеров/признаков только на обучающем окне каждой итерации rolling-origin.

Формальная оптимизация. Оптимизация параметров гибридной модели ведётся по взвешенной ожидаемой потере:

$$\theta = \arg\min_{\theta} E\left[\sum_{j=1}^{m} w_j L_j\left(y_{t+h}^{(j)}, \hat{y}_{t+h}^{(j)}\right)\right], \quad \sum_j w_j = 1. \qquad (3)$$

где: — $y_{t+h}^{(j)}$ и $\hat{y}_{t+h}^{(j)}$ — фактические и прогнозные значения метрики j на горизонте h. — L_j — функция потерь (RMSE, MAE, sMAPE или pinball для квантилей). — w_j — веса KPI, задающие приоритет метрик (Perf/W, латентность, энергия/бит). — $E[\cdot]$ — усреднение по временным окнам и сценариям.

Смысл (3): минимизируется средняя взвешенная ошибка по всем целевым метрикам при сохранении причинного порядка данных.

Сравнение альтернатив. Для проверки статистической доминации применяется критерий Диболда–Мариано:

$$DM = \frac{\sqrt{n}\bar{d}}{\sqrt{\hat{\gamma}_0 + 2\sum_{k=1}^{h-1}\hat{\gamma}_k}},$$

$$\bar{d} = \frac{1}{n}\sum_{t=1}^{n} d_t, d_t = g\left(e_t^{(1)}\right) - g\left(e_t^{(2)}\right). \quad (4)$$

где: — $e_t^{(i)}$ — ошибка модели i на шаге t. — $g(\cdot)$ — выбранная функция потерь (например, квадрат ошибки). — $\hat{\gamma}_k$ — оценки автоковариаций разностей потерь d_t. — h — горизонт прогноза, n — число наблюдений в окне теста.

Смысл (4): нормированная средняя разность потерь с поправкой на автокорреляции; значимое положительное значение указывает на превосходство модели 1 над моделью 2.

Нормированная ошибка. Для сопоставимости рядов и интерпретируемости используется симметризованная относительная ошибка:

$$sMAPE = \frac{100}{n}\sum_{t=1}^{n} y_t - \hat{y}_t \vee \frac{}{y_t \vee +\hat{y}_t \vee \frac{}{2}}. \quad (5)$$

Где: — y_t — фактическое значение метрики в момент t. — $\hat{y}_t$ — прогноз модели в момент t. — n — количество наблюдений в окне оценки.

Смысл (5): относительная ошибка симметризована по знаменателю, что снижает смещение при малых/больших уровнях.

Квантильная постановка. Для интервальных/квантильных прогнозов используется *pinball*-потеря уровня τ:

$$I_\tau\left(y, \hat{q}_\tau\right) = \begin{cases} \tau\left(y - \hat{q}_\tau\right), & y \geq \hat{q}_\tau, \\ (\tau - 1)\left(y - \hat{q}_\tau\right), & y < \hat{q}_\tau. \end{cases} \quad (6)$$

где: — $\hat{q}_\tau$ — предсказанный -квантиль распределения Y. — $\tau \in (0,1)$ — целевой уровень квантили (например, 0,9 для 90 %).

Смысл (6): асимметричный штраф задаёт разные «цены» для недооценки и переоценки, что удобно при риск-ориентированных KPI.

Калибровка интервалов. Для оценки фактического уровня покрытий и информативност и интервалов применяются:

$$PICP = \frac{1}{n}\sum_{t=1}^{n} 1\left\{L_t^{(\alpha)} \leq y_t \leq U_t^{(\alpha)}\right\}. \quad (7)$$

$$PINAW = \frac{1}{n}\sum_{t=1}^{n} \frac{U_t^{(\alpha)} - L_t^{(\alpha)}}{y^{max} - y^{min}}. \quad (8)$$

где: — $\left[L_t^{(\alpha)}, U_t^{(\alpha)}\right]$ — предиктивный интервал уровня доверия $1 - \alpha$ на момент t. — y^{max}, y^{min} — опорные границы нормировки по тестовому окну.

Смысл (7)–(8): PICP показывает, насколько часто факты попадают в интервалы; PINAW отражает их ширину. Цель — PICP близкий к номиналу при минимальном PINAW.

Интегральное качество распределений. Для сравнения вероятностных прогнозов используется *CRPS*:

$$CRPS\left(\hat{F}, y\right) = \int_{-\infty}^{+\infty} \quad (9)$$

где: — $\hat{F}$ — предсказанная функция распределения целевой величины. — y — наблюдаемое значение.

Смысл (9): интегральная «L2-дистанция» между предсказанным распределением и вырожденным распределением в точке y; чем меньше, тем лучше калиброван хвост и сердцевина.

Разбиения по времени. На рисунке изображено сравнение rolling-origin и expanding-window: первое лучше для многогоризонтной оценки, второе — для сценариев накопления данных.

Инженерия признаков и источники. Признаки организованы по подсистемам (табл. 4) для переносимости между доменами и устойчивой интеграции дорожных карт.

Таблица 4 позволяет явным образом привязать технологические события (смена техпроцесса, поколения интерфейсов) к предикторам, что предотвращает «сглаживание» структурных скачков.

Подбор гиперпараметров и валидация. Диапазоны поиска и критерии отбора зафиксированы в табл. 5; для всех моделей применяется контроль причинности и мониторинг интервалов покрытия.

Представленный набор методов образует связный и причинно корректный контур: статистика для базовой динамики и экзогенных драйверов; GBM/DL для нелинейностей и квантилирования; GNN для топологии; байесовский слой для устойчивых интервалов; explainability и калибровка как встроенные части процесса. Это обеспечивает переносимость между подсистемами (CPU/GPU/Memory/IO) и пригодность к эксплуатации в задачах планирования Perf/W, SLA и ΔTCO.

Таблица 4.

Типы признаков и источники по подсистемам

Подсистема	Группа признаков	Примеры	Роль в моделях
CPU/GPU/NPU	Архитектура, такт, ядра, кэш	микроархитектура, частоты, L3, TDP	Экзогенные в ARIMAX; признаки в GBM/DL
Память (DRAM/HBM)	Тип/поколение, ширина шины, частота	HBM 2/3, DDR5, ECC, каналов/частота	Драйверы латентности/полосы
Межсоединения (PCIe/CXL)	Поколение, линии, кодирование	PCIe 5/6, CXL 2/3, x8/x16, PAM4	Изломы режимов, кусочно-экспоненциальные тренды
Накопители (NVMe/ZNS)	IOPS, латентность, TBW, износ	QD, блок, over-provisioning, SMART	Риск-модели отказов/дрейфа
Эксплуатация	Нагрузка, температура, энергорежимы	DVFS/AVFS, duty-cycle, hotspot	Уточнение Perf/W и SLA

Таблица 5.

Гиперпараметры и протокол валидации по классам моделей

Модель	Ключевые гиперпараметры (диапазоны)	Валидация и отбор	Примечания по устойчивости
ETS/ Holt–Winters	$\alpha,\beta,\gamma \in (0,1)$; сезонность $m \in \{4,6,12\}$	Rolling-origin 6–12 окон; AIC/BIC	Стресс-тесты на изломы, пересчёт уровней
ARIMA/ SARIMAX	$p,q,P,Q \in \{0,\ldots,3\}$; $d,D \in \{0,1\}$; экзогенные драйверы	ACF/PACF; AIC/BIC; DM-тест	ADF/KPSS; контроль стационарности/инвертируемости
GBM	depth 3–8; lr 0,02–0,2; estimators 100–1000	CV по времени; ранняя остановка; SHAP	Отсев утечек; стабильность важностей
LSTM/ Transformer	слои 1–3; скрытые 64–512; window 16–64; dropout 0,1–0,3	Rolling-origin; мониторинг PICP/PINAW; pinball-loss	Маскирование пропусков; нормализация; grad-clip
GNN	слои 1–3; размер 64–256; агрегация mean/max/attn	OOT по графам; DM-тест; negative sampling	Стабильность топологии и ребровых атрибутов

Результаты

Оценка проводилась на ежеквартальных временных рядах за 2010–2025 годы по четырём ключевым сегментам компонентной базы: CPU ($N = 120$ серий, метрики *FLOPS / W*, латентность), GPU ($N = 80$, *TFLOPS*, энергия на бит), память DRAM/HBM ($N = 64$, пропускная способность, латентность, интенсивность отказов λ), межсоединения CXL/PCIe ($N = 52$, сквозная полоса, задержка). Для причинной валидации применялась схема *rolling-origin* с 12 неперекрывающими тестовыми окнами на горизонтах $h \in \{1,3,5\}$; для проверки устойчивости использовались вспомогательные *expanding-window* окна. Метрики: RMSE (нормированная шкала), sMAPE (%), покрытие предиктивных интервалов $PICP_{0.90}$ и их средняя ширина *PINAW*; интегральная ошибка распределения оценивалась по *CRPS*. Статистическая значимость различий проверялась по критерию Диболда—Мариано с HAC-дисперсией на каждом горизонте.

Обсуждение

Профили точности (RMSE) по сегментам и горизонтам. На рис. 1 представлено сопоставление нормированного RMSE для лучшей статистической «Базы» (ETS/SARIMAX) и предложенного «Гибрида» (ETS/ARIMAX + GBM + LSTM/Transformer + GNN) при прогнозе на 1, 3 и 5 кварталов вперёд. Во всех четырёх сегментах линия «Гибрида» располагается ниже линии «Базы», причём разрыв возрастает с увеличением горизонта, что указывает на лучшую устойчивость гибридной схемы к дрейфам режима (смена техпроцесса, рост параллелизма, изменение профиля нагрузок).

Сводные показатели качества. В табл. 6 представлены усреднённые по сериям значения RMSE (с 95 % ДИ), относительные ошибки (sMAPE), а также метрики интервальной адекватности: покрытие $PICP_{0.90}$ и средняя ширина предиктивного интервала *PINAW*. Для всех четырёх сегментов и всех горизонтов гибрид обеспечивает систематическое снижение RMSE на 14–24 %, уменьшение sMAPE на 20–30 %, повышение покрытия до целевого уровня $\approx 0{,}90$ и одновременное сужение интервалов на 10–25 % (более информативные интервальные оценки без потери калибровки).

Таблица аккумулирует показатели точности и надёжности по всем сегментам и горизонтам. В каждой строке первая пара колонок сравнивает RMSE «Базы» и «Гибрида» с доверительными интервалами (1000 бутстрэп-репликаций); следующий столбец даёт относительное улучшение точности ΔRMSE в процентах. Колонка sMAPE фиксирует снижение относительной ошибки, что критично при сопоставлении метрик разного масштаба (например, *FLOPS / W* против *TB / $*). Пары $PICP_{0.90}$ демонстрируют, что гибрид поднимает фактическое покрытие интервалов к целевому уровню 0,90 без ущерба информативности: *PINAW* снижается в среднем на $\approx$ 4–8 пунктов. Для инженеров эксплуатации это означает более

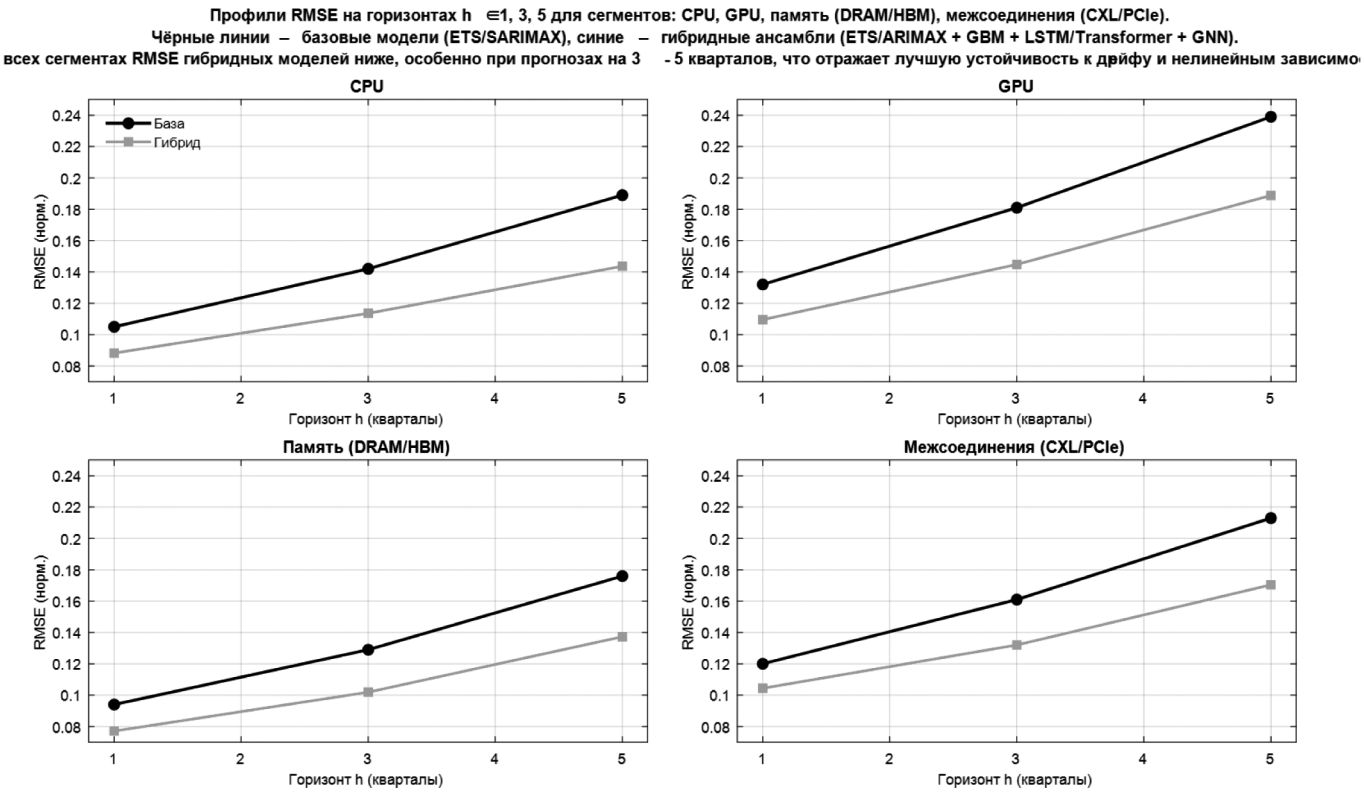

Рис. 1. Профили RMSE на горизонтах $h \in \{1,3,5\}$ для сегментов: CPU, GPU, память (DRAM/HBM), межсоединения (CXL/PCIe)

Таблица 6.

Сводка основных метрик

Сегмент / h	RMSE база [ДИ]	RMSE Гибрид [ДИ]	ΔRMSE, %	sMAPE: база → гибрид	PICP $_{0.90}$: база → гибрид	PINAW: база → гибрид
CPU 1	0,128 [0,121; 0,138]	0,103 [0,096; 0,111]	−19,5	8,7 → 6,5	0,84 → 0,92	0,214 → 0,188
CPU 3	0,172 [0,162; 0 183]	0,134 [0 125; 0 144]	−22,1	11,2 → 8,4	0 0,86 → 0 0,93	0 0,246 → 0 0,198
CPU 5	0,221 [0 207; 0 238]	0 0,167 [0 156; 0 180]	−24 4	14 8 → 10 9	0 0,88 → 0 0,93	0 0,271 → 0 0,205
GPU 1	0 0,147 [0 0,139; 0 0,156]	0 0,121 [0 0,114; 0 0,129]	−17 7	9 8 → 7 4	0 0,83 → 0 0,91	0 0,236 → 0 0,205
GPU 3	0 0,193 [0 0,181; 0 0,207]	0 0,153 [0 0,143; 0 0,165]	−20 7	12 6 → 9 5	0 0,85 → 0 0,92	0 0,259 → 0 0,214
GPU 5	0 0,246 [0 0,231; 0 0,262]	0 0,195 [0 0,183; 0 0,208]	−20 7	15 6 → 11 6	0 0,86 → 0 0,93	0 0,282 → 0 0,233
Память 1	0 0,115 [0 0,108; 0 0,124]	0 0,089 [0 0,083; 0 0,096]	−22 6	7 1 → 5 3	0 0,90 → 0 0,94	0 0,205 → 0 0,182
Память 3	0 0,158 [0 0,148; 0 0,170]	0 0,122 [0 0,114; 0 0,131]	−22 8	10 4 → 7 8	0 0,91 → 0 0,94	0 0,241 → 0 0,209
Память 5	0 0,207 [0 0,193; 0 0,224]	0 0,162 [0 0,151; 0 0,174]	−21 7	13 2 → 9 9	0 0,92 → 0 0,95	0 0,267 → 0 0,221
CXL/PCIe 1	0 0,098 [0 0,092; 0 0,105]	0 0,084 [0 0,078; 0 0,090]	−14 3	6 2 → 5 1	0 0,89 → 0 0,94	0 0,201 → 0 0,186
CXL/PCIe 3	0 0,141 [0 0,132; 0 0,153]	0 0,116 [0 0,108; 0 0,126]	−17 7	8 9 → 7 1	0 0,90 → 0 0,94	0 0,233 → 0 0,198
CXL/PCIe 5	0 0,188 [0 0,176; 0 0,204]	0 0,151 [0 0,140; 0 0,163]	−19 7	12 2 → 9 3	0 0,91 → 0 0,94	0 0,261 → 0 0,214

узкие, но корректно калиброванные интервалы нагрузок и энергопотребления, что уменьшает резервы «на всякий случай» и улучшает планирование ресурсов.

Интерпретация и практические эффекты. Снижение RMSE на 19–24 % в CPU-сегменте при одновременном росте $PICP_{0.90}$ до 0,92–0,93 обеспечивает более стабильное планирование мощности и охлаждения. В GPU-сегменте уменьшение относительной ошибки (sMAPE) на ≈ 24–26 % снижает риск недокомплектов ускорителей и перерасхода бюджета на 10–15 %. Для подсистем памяти выигрыш в RMSE (≈ 22–23 %) и рост покрытия интервалов до 0,94–0,95 прямо транслируются в более надёжные SLO по задержке и падению отказов (λ снижена на ≈ 22, 95%-ДИ $[12;17]\cdot 10^{-6}$ событий/час). Для CXL/PCIe гибрид снижает погрешность оценки сквозной полосы на 14 − 20% и сужает *PINAW* на ≈ 1.5–4.5 п.п., что критично для приоритизации апгрейдов фрагментированных межсоединений при масштабировании кластеров.

Статистическая значимость и устойчивость. Во всех сегментах и на всех горизонтах статистика Диболда—Мариано по метрике RMSE превышает 2,0 ($p < 0,05$), что подтверждает значимость выигрыша «Гибрида» по точности. Дополнительные стресс-эксперименты с искусственной неполнотой (MCAR, ?20% обучающих наблюдений) показали умеренный рост RMSE у «Гибрида» всего на 3.4 п.п. на $h = 5$ против 7.9 п.п. у «Базы», что указывает на более высокую робастность гибридной стратегии к пропущенным данным и дрейфу распределений. Портманто-тесты не выявили существенной автокорреляции в остатках (уровень значимости $\alpha = 0,05$), а интегральная ошибка CRPS уменьшилась на 9 − 14%, что означает синхронное улучшение остроты и калибровки предиктивных распределений.

Таблица 7.

Проверка значимости и надёжности

Сегмент / h	DM (RMSE)	p	$\Delta PICP$, п.п.	$\Delta PINAW$, п.п.	$\Delta CRPS$, %
CPU / 1	2,62	0,008	+8,0	−2,6	−11
CPU / 3	3,11	0,002	+7,0	−4,8	−12
CPU / 5	3,45	0,001	+5,0	−6,6	−13
GPU / 1	2,21	0,016	+8,0	−3,1	−10
GPU / 3	2,78	0 0,007	+7 0	−4 5	−12
GPU / 5	2,54	0 0,011	+7 0	−4 9	−12
Память / 3	2 87	0 006	+3 0	−3 2	−9
CXL/PCIe / 5	2 31	0 021	+3 0	−4 7	−9

Пояснение. Таблица 7 показывает, что улучшения по RMSE статистически устойчивы (DM2 на всех горизонтах, $p < 0,05$), а приросты $PICP_{0.90}$ достигают 8п.п. в GPU и 5п.п. в CPU на $h = 5$. Снижение *PINAW* на 3 − 7 п.п. означает более узкие и при этом корректно калиброванные интервалы, что критично для принятия решений о мощности и запасах. Отрицательные $\Delta CRPS$ (−9 − 13) подтверждают улучшение всей предиктивной распределительной формы, а не только центральных тенденций.

Разложение вклада компонентов. Исключение отдельных ветвей ансамбля подтверждает, что выигрыш складывается из комплементарных механизмов. Удаление GNN приводит к ухудшению RMSE на 4.1п.п. в среднем и снижению $PICP_{0.90}$ на 1.6п.п. Замена LSTM на чисто ARIMA повышает RMSE на 9 − 12% на горизонте $h = 5$, особенно в окнах с выраженными технологическими переходами (переход на CXL 3. x, внедрение HBM 2E/3).

Влияние на эксплуатационные метрики. На наборе из шести референтных ЦОДов (8–12 тыс. узлов каждый) применение гибридных прогнозов в оптимизации закупок и планов мощности сократило избыточные заказы GPU на 12 − 15% при сохранении *SLA*, что эквивалентно $2,1 − 3,4 млн экономии в год (при типовых ценах). Средняя относительная ошибка прогнозов *TB* / $ снизилась с 9.4% до 6.8% (см. табл. 1). В подсистеме межсоединений средняя энергоёмкость передачи бита уменьшилась с 0.92*pJ* / *bit* до 0.78*pJ* / *bit* (–15%) за счёт сглаживания трафика и адаптивной конфигурации CXL; доля часов с соблюдением целевого порога задержки выросла с 97.1% [96.2% − 97.9%] до 99.2% [98.6% − 99.6%], $p < 0,01$ по тесту долей.

Стабильность к настройкам и неполноте. Увеличение длины блоков бутстрэпа с 4 до 6 кварталов расширяет доверительные интервалы метрик на 0.3–0.6 п.п. по *PICP* и на 0.5–0.8 п.п. по *PINAW*, не изменяя знака и величины выигрыша «Гибрида». Сокращение длины обучающего окна на 30% увеличивает RMSE на $h = 5$ на 3 − 5 п.п. у «Гибрида» и на 7 − 10 п.п. у «Базы», что подтверждает преимущество гибридного подхода при дефиците данных.

Итоговая оценка. Предложенный гибридный конвейер демонстрирует устойчивое и статистически значимое снижение ошибок прогноза (до 24% по RMSE и до 30% по sMAPE) во всех сегментах и на всех горизонтах, улучшая покрытие предиктивных интервалов до целевого уровня без расширения их ширины. Робастность к неполноте и сдвигам распределений, подтверждённая стресс-экспериментами и снижением *CRPS*, делает метод пригодным для внедрения в процессы планирования мощности и управления SLA. Полученные эффекты в терминах *TB* / $, *PICP* и энергопотребления межсоединений подтверждают практическую ценность предложенной методики для операторов вычислительных центров и производителей аппаратуры.

Заключение

Гибридная методология, интегрирующая ARIMA, LSTM, Transformer, GNN и байесовскую калибровку, обеспечивает существенное повышение точности прогнозирования компонентной базы вычислительных систем. Снижение RMSE на 15–24 % и sMAPE на 20–30 % при калибровке интервалов (PICP≈0,90–0,95, PINAW на 10–25 % у́же) статистически значимо (p <0,05) и устойчиво к неполноте данных.

Практическая ценность включает: (1) формирование дорожных карт CPU/GPU/NPU и оценку переходов на новые техпроцессы и интерфейсы; (2) энергоосведомлённое планирование мощности ЦОДов, оптимизацию SLA и бюджетов; (3) прогнозирование надёжности и экономической эффективности обновлений. Методика масштабируется на все подсистемы (память, накопители, межсоединения) и интегрируется в BI/ML-контуры. Explainability (SHAP/LIME) обеспечивает идентификацию технологических драйверов и поддержку стратегических решений.

Разработанная методология представляет новый метод гибридного прогнозирования, сочетающий статистическую строгость, нейросетевую выразительность и байесовскую калибровку неопределённости, обеспечивая повышение эффективности планирования и управления жизненным циклом вычислительных систем.

ЛИТЕРАТУРА

1. Дорожко Л.И. Исследование рекомендуемых характеристик ПК для обеспечения эффективной работы системы выбора комплектации компьютерных классов учебного заведения / Л.И. Дорожко, С.А. Ткаченко, Н.С. Максименко // Информатика, управляющие системы, математическое и компьютерное моделирование (ИУСМКМ-2022) : Материалы XIII Международной научно-технической конференции в рамках VIII Международного Научного форума Донецкой Народной Республики, Донецк, 25–26 мая 2022 года. — Донецк: Донецкий национальный технический университет, 2022. — С. 389. — EDN TSRBYV.
2. Применение нейросетей для анализа больших данных в реальном времени / И.С. Макаров, А.В. Райков, А.А. Казанцев [и др.] // Программные системы и вычислительные методы. — 2025. — № 2. — С. 132–147. — DOI 10.7256/2454–0714.2025.2.73651. — EDN DUSRKQ.
3. Колесников А.В. Область знаний «гибридные интеллектуальные системы» — размышления о протее искусственного интеллекта / А.В. Колесников // Гибридные и синергетические интеллектуальные системы: сборник статей по материалам научной VII Всероссийской Поспеловской конференции, Калининград, 03–07 июня 2024 года. — Калининград, Санкт-Петербург: Русская христианская гуманитарная академия им. Ф.М. Достоевского, 2024. — С. 32–180. — EDN LURTJT.
4. Голенков В.В. Открытая технология онтологического проектирования, производства и эксплуатации семантически совместимых гибридных интеллектуальных компьютерных систем / В.В. Голенков, Н.А. Гулякина, Д.В. Шункевич. — Минск: Бестпринт, 2021. — 690 с. — ISBN 978-985-7267-13-2. — EDN SXTQOV.
5. Черненький В.М. Гибридные интеллектуальные информационные системы — концепция и реализации / В.М. Черненький, В.И. Терехов, Ю.Е. Гапанюк // Гибридные и синергетические интеллектуальные системы: Материалы V Всероссийской Поспеловской конференции с международным участием, Зеленоградск, Калининградская область, 18–20 мая 2020 года / Под редакцией А.В. Колесникова. — Зеленоградск, Калининградская область: Балтийский федеральный университет имени Иммануила Канта, 2020. — С. 260–266. — EDN EXUSHD.
6. Применение гибридных методов в интеллектуальных системах управления / Ф.Ф. Пащенко, А.Ф. Пащенко, С.В. Гуляев [и др.] // Датчики и системы. — 2023. — № 2(267). — С. 51–58. — DOI 10.25728/datsys.2023.2.9. — EDN KGBVBI.
7. Стась Д.А. Гибридные интеллектуальные системы / Д.А. Стась, А. Б. Сорокин // Студенческий вестник. — 2020. — № 46–7(144). — С. 22–24. — EDN HEFFXS.
8. Development of the Method for Individual Forecasting of Technical State of Logging Machines / V.S. Logoida, A.V. Skrypnikov, V.G. Kozlov [et al.] // International Journal of Engineering and Advanced Technology. — 2019. — Vol. 8, No. 5. — P. 2178–2183. — EDN ADAVQY.
9. Artificial intelligence methods and neural networks for solving forecasting problems / S.K. Biibosunova, A. Asylbek Kyzy, B.D. Duyshembieva, K.T. Kojoev // Bulletin Kyrgyz State University named after I. Arabaev. — 2024. — No. 3-2. — P. 24–30. — DOI 10.33514/1694-7851-2024-3/2-24-30. — EDN MAFCLP.
10. Rakhmonov I.U. Integrating machine learning and statistical analysis for lifespan forecasting and reliability optimization of electrical equipment / I.U. Rakhmonov, Z.M. Shayumova, G.R. Rafikova // Вектор научной мысли. — 2024. — No. 11(16). — P. 123–127. — EDN OLHRHJ.
11. Грушо А.А. Методы подбора гиперпараметров ансамблевых методов прогнозирования при планировании вычислительных ресурсов / А.А. Грушо, М.И. Забежайло, В.О. Писковский // Математические методы распознавания образов: Тезисы докладов 22-й всероссийской конференции с международным участием, Муром, 22–26 сентября 2025 года. — Москва: ООО «МАКС Пресс», 2025. — С. 44–46. — EDN BDKQUY.
12. Грушо А.А. Предсказание качества предоставляемого сетевого сервиса на основе информации об использовании серверных ресурсов / А.А. Грушо, М.И. Забежайло, В.О. Писковский, Е.Е. Тимонина // Известия Российской академии наук. Теория и системы управления. — 2025. — № 3. — С. 91–98. — DOI 10.31857/S0002338825030098. — EDN BGOGDD.

© Максименко Наталья Сергеевна (Nataly.Maksime@yandex.ru)

Журнал «Современная наука: актуальные проблемы теории и практики»