УДК 621.396.06 Научная статья

https://doi.org/10.32603/2071-8985-2025-18-9-56-67

Параллельный метод ассемблирования нейронных сетей в подсистеме прогнозирования неисправностей высокопроизводительных вычислительных комплексов

В. Ю. Мельцов, А. К. Крутиков✉

Вятский государственный университет, Киров, Россия

✉ usr09603@vyatsu.ru

Аннотация. Рассматриваются вопросы повышения эффективности применения искусственных нейронных сетей для прогнозирования сбоев и отказов высокопроизводительных вычислительных комплексов в реальном времени. Особое внимание уделяется многокомпонентным системам со сложной коммутацией и массовым параллелизмом. Для повышения точности прогнозирования сбоев и отказов предлагается параллельный метод ассемблирования нейронных сетей на основе выявления внутренних логических взаимосвязей между различными элементами комплекса и выполнения специальной фрагментации обучающих выборок для каждого яруса. Синтезирован экспериментальный комплекс на базе кластерной системы и проведено сравнительное тестирование различных конфигураций модуля прогнозирования. Задача своевременной диагностики и предсказания нарушений работоспособности специализированных цифровых комплексов приобретает особую значимость, так как серьезные инциденты способны вызвать не только приостановку функционирования аппаратуры, но и утрату критически важных данных.

Ключевые слова: вычислительный комплекс, кластерная система, диагностика, прогнозирование отказов, искусственная нейронная сеть, фрагментированная выборка, каскадная архитектура, журнал событий, программный прототип

Для цитирования: Мельцов В. Ю., Крутиков А. К. Параллельный метод ассемблирования нейронных сетей в подсистеме прогнозирования неисправностей высокопроизводительных вычислительных комплексов // Изв. СПбГЭТУ «ЛЭТИ». 2025. Т. 18, № 9. С. 56–67. doi: 10.32603/2071-8985-2025-18-9-56-67.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Original article

Parallel Method for Assembling Neural Networks in a Fault Prediction Subsystem of High-Performance Computing Complexes

V. Yu. Meltsov, A. K. Krutikov✉

Vyatka State University, Kirov, Russia

✉ usr09603@vyatsu.ru

Abstract. The article discusses the issues of increasing the efficiency of using artificial neural networks to predict failures and failures of high-performance computing complexes in real time. Particular attention is paid to multicomponent systems with complex switching and massive parallelism. To improve the accuracy of forecasting failures and breakdowns, a parallel method of assembling neural networks is proposed based on identifying internal logical relationships between different elements of the complex and performing special fragmentation of training samples for each tier. An experimental complex based on a cluster system is synthesized and comparative testing of the forecasting module various configurations is carried out. The task of timely diagnostics and prediction of specialized digital complexes malfunctions is of particular importance, since serious incidents can cause not only a suspension of equipment operation, but also the loss of critical data.

Keywords: computing complex, cluster system, diagnostics, failure prediction, artificial neural network, fragmented sampling, cascade architecture, event log, software prototype

For citation: Meltsov V. Yu., Krutikov A. K. Parallel Method for Assembling Neural Networks in a Fault Prediction Subsystem of High-Performance Computing Complexes // LETI Transactions on Electrical Engineering & Computer Science. 2025. Vol. 18, no. 9. P. 56–67. doi: 10.32603/2071-8985-2025-18-9-56-67.

Conflict of interest. The authors declare no conflicts of interest.

Введение. В настоящее время информационные технологии занимают важнейшее положение в динамичном развитии как промышленного производства, так и социальной сферы. Среди многообразия инновационных подходов особо выделяются два направления: прикладной искусственный интеллект и высокопроизводительные вычислительные системы (ВПВС), в английской нотации – High Performance Computers (HPC). Последние представляют собой мощнейшие супер-ЭВМ, ориентированные на решение комплексных проблем, характеризуемые высокой потребностью в интенсивных расчетах и аналитической обработке огромных массивов данных. Применение высокопроизводительных вычислений охватывает широкий спектр областей человеческой деятельности, начиная от фундаментальных научных исследований, моделирования искусственных нейронных сетей, изучения климатических изменений и заканчивая глубоким анализом финансово-экономических показателей. В последнее время именно возрастание быстродействия супер-ЭВМ способствует резкому «повышению интеллекта» автоматизированных систем обработки информации и специализированных вычислительных комплексов.

Однако суперкомпьютеры, подобно любым многокомпонентным техническим системам, сталкиваются с риском сбоев и функциональных отказов, обусловленных сложностью их конструкции и высокими нагрузками при эксплуатации. Процесс диагностического мониторинга цифровых систем с массовым параллелизмом направлен на предупреждение потенциальных аварийных состояний и снижение вероятности сбоев и отказов оборудования. К сожалению, данный процесс становится крайне трудоемким с учетом продолжающегося роста аппаратурных масштабов HPC. Число конструктивных элементов обработки информации (элементарных машин – ЭМ) в них уже сейчас имеет порядок $10^7 \ldots 10^8$. Серьезные инциденты способны вызвать приостановку функционирования аппаратуры, утрату критически важных данных и значительное ухудшение эксплуатационных характеристик. Вследствие этого задача своевременной диагностики и предсказания нарушений работоспособности подобных комплексов приобретает особую значимость и требует повышенного внимания специалистов.

Существующие методы диагностики и прогнозирования. Традиционно для прогнозирования и диагностики состояния цифровых устройств и обеспечения возможности корректной обработки данных используется классический аппарат математической статистики [1]–[3]. Методы статистического анализа – хороший инструмент для изучения распределения значений различных метрологических параметров, оценки средней продолжительности жизни компонентов, определения частоты возникновения типичных ошибок и прогнозирования сроков их проявления [4]. Применение регрессионного анализа позволяет строить модели зависимостей между исходными факторами и конечными показателями работоспособности, что облегчает раннюю диагностику возникающих неполадок [5], [6]. Вероятностные методы помогают рассчитывать степень надежности каждого компонента системы и формировать рекомендации по проведению профилактического ремонта или замены деталей, достигших предельного ресурса [7], [8]. Основополагающим элементом данных систем служит упомянутый ранее инструментальный аппарат математической статистики. Использование подобных методов в цифровых технологиях обязательно должно учитывать все факторы, оказывающие влияние на рабочие показатели. Такими факторами могут выступать: специализированные технические параметры самого оборудования, объем запросов и активность пользователей, уровень загруженности тех или иных модульных сегментов и т. п. [9]. Особое значение среди статистических методов имеет техника корреляционного анализа, позволяющая устанавливать связи и взаимодействия между функционированием вычислительных элементов и сопутствующими внешними условиями, например характером нагрузок или спецификой аппаратного окружения [10]–[12].

Однако применение алгоритмов математической статистики в процессе прогнозирования и диагностики сложных вычислительных комплексов с массовым параллелизмом обладает рядом ограничений и недостатков. Статистические методы основаны преимущественно на анализе прошлых данных, что затрудняет учет влияния случайных факторов и внешних воздействий, приводящих к непредвиденным изменениям состояния системы. А рост количества и разнообразия подобных факторов в современных HPC происходит постоянно. Оценка надежности компонентов методами теории вероятностей предполагает идеальные условия эксплуатации, игнорируя реальные факторы старения, деградации материалов и, особенно, резкое изменение температурных режимов в плотной компоновке стоек современных кластеров. Более того, традиционные статистические подходы не учитывают сложность современных многоуровневых архитектур вычислительных систем, где локальная ошибка может иметь каскадный эффект, распространяющийся на всю инфраструктуру. Все это делает использование исключительно статистических инструментов недостаточным для полноценной диагностики и прогнозирования в области высокопроизводительных аппаратно-программных комплексов.

Для решения вышеуказанных проблем сегодня все чаше применяют новые, более современные методы планирования и прогнозирования [13]. Среди высокоэффективных механизмов прогнозирования важное место занимает технология искусственных нейронных сетей (ИНС), получившая широкое распространение благодаря своим уникальным свойствам и возможностям адаптивного машинного обучения [14], [15]. Используя способности к обучению и самообучению, гибкости настройки, ИНС демонстрируют высокую эффективность в решении сложных задач диагностики работоспособности масштабных (massively) цифровых систем. Вместе с тем, внедрение нейронных сетей предъявляет определенные требования к объему анализируемых данных, а также качеству предварительной обработки и нормализации обучающей выборки (ОВ) [16].

Методология и методы исследований. Как уже было сказано, супер-ЭВМ представляет собой сложную аппаратно-программную систему, включающую совокупность взаимосвязанных вычислительных узлов различной архитектуры и функционального назначения. Поломка отдельного компонента, и даже отдельного логического элемента, способна существенно ухудшить общую производительность системы вплоть до полной остановки ее функционирования. Особенно это критично в условиях жесткой взаимозависимости элементов, определяемой конкретной коммуникационной топологией. Высокопроизводительный вычислительный комплекс, в сущности, есть виртуальная система или, точнее, программно настроенная конфигурация, обладающая следующими особенностями [17]:

1. Выделены основная подсистема из n элементарных машин (ЭМ) и подсистемы, составляющие избыточность из $(N - n)$ машин ($n \neq 0$, $n \in E^N$), где E^N – множество состояний элементарных машин.

2. Основная подсистема предназначена для решения сложных задач из n ветвей, а любая подчиненная подсистема – для решения фоновых задач.

3. Функции отказавшей ЭМ основной подсистемы может взять на себя любая исправная ЭМ любой подчиненной подсистемы.

Тогда функцию надежности можно определить как вероятность того, что производительность P вычислительной системы (ВС), начавшей функционировать в состоянии i ($n \le i \le N$), на промежутке времени $[0, t)$ равна производительности основной подсистемы:

$$R(t) = P\{\forall \tau \in [0,1] \ \rightarrow \Omega(\tau) \ = A_n \omega_n \mid n \le i \le N\},$$

где $\Omega(\tau)$ – производительность системы в момент времени τ; A_n – специальный показатель алгоритма взаимодействий элементарных машин; ω – среднее число операций, выполняемых ЭМ в секунду.

Говоря иначе, функция $R(t)$ есть вероятность того, что в системе на промежутке времени $[0, t)$ будет не менее n исправных машин.

Под методологией прогнозирования будем понимать объективно обоснованное научное предвидение будущих событий на основании ретроспективного анализа текущих и исторических тенденций. Данный подход широко применяется во многих отраслях научного знания и промышленной практики. Применительно к диагностике компьютерных систем прогнозирование потенциального износа или выхода из строя отдельных комплектующих, или всей системы в целом играет решающую роль в обеспечении надежного функционирования технических устройств посредством заблаговременного планирования профилактических мероприятий и предотвращения появления дефектов. Важно отметить, что формальная постановка задачи прогнозирования осуществляется в двух аспектах:

– с одной стороны, как задача аппроксимации, предполагающая количественное оценивание конкретных цифровых метрик поведения вычислительного модуля или системы в целом;

– с другой стороны, как задача классификации, заключающаяся в установлении факта наступления определенного события (например, полного отказа устройства).

Оба указанных подхода дополняют друг друга и служат основой для формирования эффективной стратегии поддержания быстродействия и стабильности массовых параллельных вычислений.

Процесс прогнозирования возможных неисправностей или нарушений нормального функционирования вычислительной системы, структурированной в виде совокупности взаимодействующих функциональных блоков (ФБ), предусматривает необходимость осуществления прогноза выхода из строя нескольких взаимосвязанных функциональных элементов, формирующих целостное пространство скоммутированных между собой компонентов высокопроизводительного вычислительного комплекса (ВВК). Задача прогнозирования состоит в определении будущих значений важнейших числовых характеристик, отражающих надежность цифровой системы с массовым параллелизмом [18].

Для расчета этих метрик будем использовать классические методики [19]:

$$R(t) \ge 1 - B^{-1}\left(1 - t^{-B}\right)$$
$$\text{при } N(\tau) \ge (1+B)\ln(\tau+1)/K;$$
$$R(t) \le 2/(1+t) \text{ при } N(\tau) \le \ln(\tau+1)/K;$$
$$R(t) \sim \mathrm{e}^{\Lambda t} \text{ при } N = \text{const},$$

где B – произвольное положительное число;

$$K = \max \kappa(\tau),\ k = \min \kappa(\tau),$$
$$\kappa(\tau) = \nu(\tau)\ln\left[\nu(\tau)/\rho(\tau)\right] +$$
$$+ [1-\nu(\tau)]\ln\left[(1-\nu(\tau))/(1-\rho(\tau))\right],$$
$$\nu(\tau) = (n-1)/N(\tau),\ \rho(\tau) = n/N(\tau),$$

для константы Λ справедливо неравенство

$$\ln\left[1-\mathrm{e}^{-KN}\right] \ge \Lambda \ge \ln\left[1-\mathrm{e}^{-kN}\right].$$

Применение классических способов расчета математических ожиданий времени безотказной работы и времени восстановления для большемасштабных ВС наталкивается на серьезные препятствия, связанные с трудоемкими и сложными вычислениями функций надежности и восстановимости. Вычисления этих функций основываются на традиционных стохастических моделях теории массового обслуживания и методах приближенных вычислений. Трудоемкость такого расчета повышается с ростом количества элементарных машин в комплексе и сложности коммутационной сети супер-ЭВМ. Соответственно, в какой-то момент масштабирования системы не удается получить адекватные аналитические формулы для отыскания числовых значений $R(t)$.

Рассмотрим случай, когда функциональные блоки ФБ_1–ФБ_4 (рис. 1) характеризуются одинаковыми наборами контролируемых параметров, регистрируемых в строго установленные равные промежутки времени. В этом случае для моделей ФБ (аналогов ЭМ) гарантируется совместимость данных и удобство их обработки существующими инструментами машинного обучения. RAID (Redundant Array of Independent Disks) – специализированный модуль памяти.

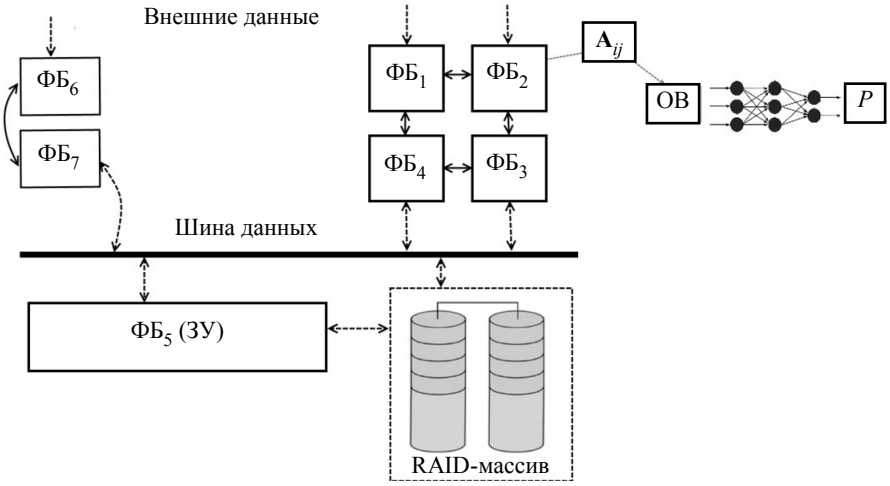

Рис. 1. Прогнозирование сбоев/отказов функциональных блоков
Fig. 1. Malfunction/failure prediction of functional blocks

Математическая интерпретация подразумевает, что совокупность всех наблюдаемых величин, фиксируемая в каждом временном интервале, образует вектор, отображающий мгновенное состояние соответствующего блока. Если рассмотреть весь период наблюдений за функционированием системы, то возможно создание обучающего датасета в форме двумерной матрицы $\mathbf{A}_{ij}$, каждая ячейка которой $a[i, j]$ показывает значение j-го параметра ФБ в i-й момент времени (табл. 1). Дополнительно формируется два отдельных одномерных вектора $\mathbf{B}_1$ и $\mathbf{B}_2$, содержащих сведения о фактах сбоев и отказов (табл. 2). Если данные события отсутствуют, значит сохраняется нормальная работоспособность блока в данный временной интервал.

Табл. 1. Фрагмент матрицы $\mathbf{A}_{ij}$
Tab. 1. Fragment of the matrix $\mathbf{A}_{ij}$

Временной интервал	Число процессов	Объем ОП, Гбайт	Рабочая частота, ГГц	Температура, °С
1	227	1.765	1.77	43
2	241	3.243	1.88	48
3	295	5.667	2.48	69
4	259	3.843	2.04	54
5	243	3.943	1.88	50
…	…	…	…	…

Табл. 2. Фрагменты векторов $\mathbf{B}_1$ (сбоев) и $\mathbf{B}_2$ (полных отказов)
Tab. 2. Fragments of $\mathbf{B}_1$ (malfunction) and $\mathbf{B}_2$ (complete failures) vectors

Временной интервал	Значение $\mathbf{B}_1$	Значение $\mathbf{B}_2$
1	0.963 (норма)	0.868 (норма)
2	1.114 (норма)	0.889 (норма)
3	1.873 (сбой)	1.932 (отказ)
4	1.611 (сбой)	0.741 (норма)
5	0.778 (норма)	0.467 (норма)

Подобная структура датасета весьма эффективна для обучения нейросетевой модели. При этом одномерные векторы будут выступать в роли векторов эталонных значений. Поскольку все ФБ однотипны, то и модели нейросети одинаковы. Выход из строя любого блока будет обозначать отказ вычислительной системы в целом.

Описанная методология моделирования показала неплохие результаты прогнозирования и применимость традиционного формата обучения определенной модели ИНС на основе обычной прямоугольной двумерной выборки лишь при условии однородности состава компонентов вычислительного комплекса. Однако ситуация кардинально меняется, если речь идет о неоднородной структуре гетерогенной вычислительной системы, включающей разнотипные блоки и модули или специализированные вычислительные узлы, отличающиеся по своему устройству и назначению. При таком сценарии возникает потребность в разработке принципиально иного подхода к формированию обучающей выборки (обучающих выборок), способного учесть разнообразие типов вычислительных компонентов и специфику сбора необходимых сведений о каждом из них. Потребуется разработка особого способа агрегации разрозненных данных, учитывающего особенности конфигурации каждой группы блоков, варьирующиеся временные интервалы регистрации показаний и прочие важные обстоятельства, усложняющие подготовку полноценных и содержательно насыщенных обучающих выборок.

Настоящее исследование предлагает формирование обучающей выборки посредством предварительного разбиения первоначального большого двумерного массива параметров на серию двумерных матриц меньшего размера – так называемое фрагментирование обучающей выборки. Основной идеей предлагаемого подхода служит разделение исходного массива на отдельные подмножества на основе наличия внутренних логических взаимосвязей между элементами данных. Логика выделения отдельных фрагментов определяется степенью взаимозависимости входящих параметров, отражающей тесноту корреляционных связей между ними и природу порождаемой информации.

Рассмотрим процесс формирования архитектурно-структурных решений интеллектуального модуля прогнозирования на основе нейросетевого подхода для специализированного гетерогенного вычислительного комплекса. Вначале синтезируются нейросетевые модули (например, на основе GRNN- или LVQ-моделей) для предсказания сбоев и/или отказов всех функциональных блоков (рис. 2).

Указанные ФБ образуют какой-либо специализированный модуль (подсистему) вычислительного комплекса. На рис. 2 продемонстрировано выделение подобного модуля, содержащего несколько разнотипных ФБ, для которого прогнозируется наступление фактов сбоев и различного типа отказов. Анализ выполняется в идентичные временные промежутки, называемые периодами наблюдения, в течение которых регистрируются соответствующие характеристики, выбранные ФБ. Эта информация последовательно заносится в специальную матрицу параметров таким образом, что каждая строка матрицы соответствует определенному моменту времени, а столбцы образуют

набор характеристик отдельных функциональных блоков. При формировании матрицы соблюдается условие принадлежности каждого отдельного параметра матрице параметров конкретного ФБ исследуемого модуля, так, что

$$A_{ij}^{f} \Rightarrow \mathbf{B}_{n}^{m},$$

где f ($f = 1, F$) – номер блока в модуле; n ($n = 1, 2$) – номер вектора в промежуточной выборке; m ($m = 1, M$) – номер модуля в системе.

Таким образом формируются строки-векторы матрицы параметров модуля для последующей обработки и прогнозирования работоспособности всего вычислительного комплекса в целом (рис. 2).

Итогом обучения становится нейросеть, способная выдавать прогнозируемые результаты в виде определенных числовых показателей, отражающих ожидаемое поведение подсистемы или вероятность наступления искомых событий (сбоев и отказов).

Процедура прогнозирования работоспособности всего вычислительного комплекса повторяет общий подход, изложенный ранее, включая стадии предварительной обработки, обучения модели и выдачи финального прогноза, позволяющего своевременно обнаружить признаки потенциального отказа или снижения производительности при функционировании системы (рис. 3). Логические взаимосвязи между элементами ОВ задают-

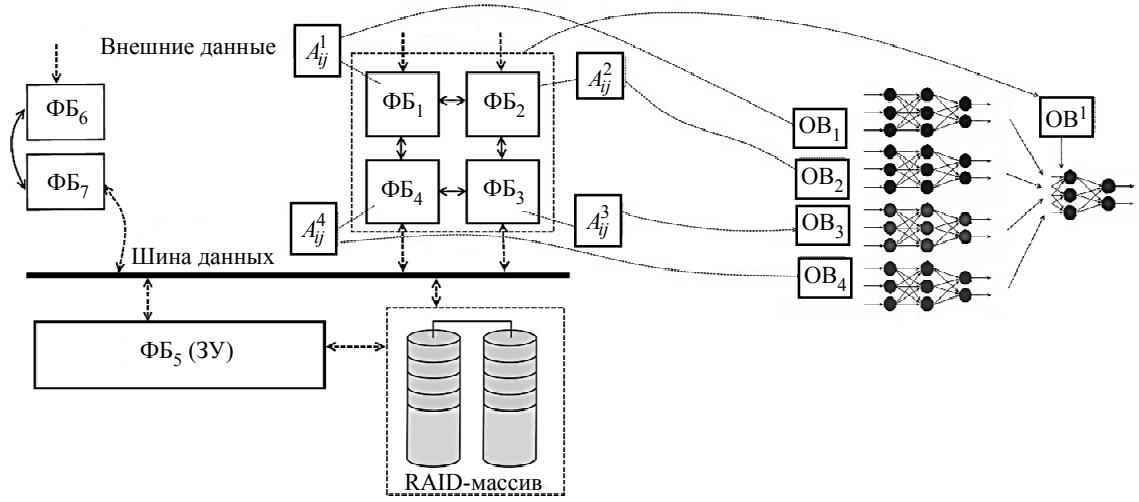

Рис. 2. Прогнозирование сбоев/отказов модуля ВВК с учетом межблоковых связей
Fig. 2. Malfunction/failure prediction of the module taking into account interblock connections

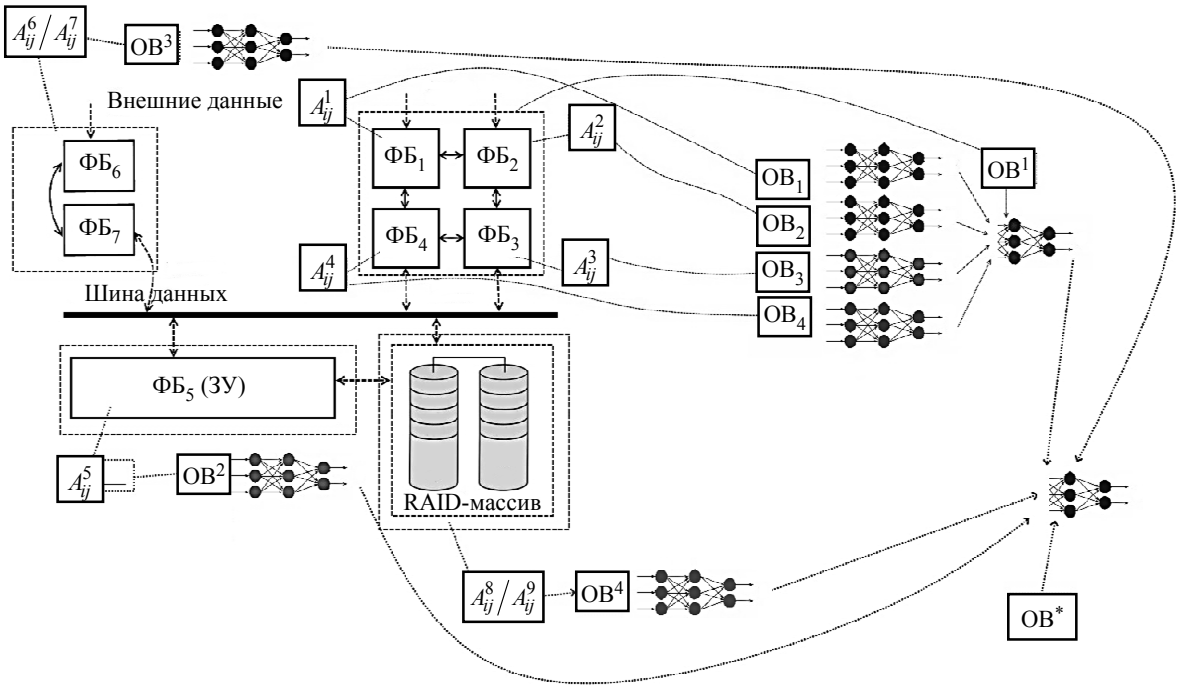

Рис. 3. Прогнозирование сбоев/отказов ВВК в целом
Fig. 3. Malfunction/failure prediction of HPC as a whole

ся исходя из внутренней топологии самой вычислительной системы, отражающей характерные свойства ее организационно-технической структуры. Выделение связей между отдельными фрагментами обучающей выборки основывается на результатах предшествующего анализа логических взаимоотношений между отдельными модулями и компонентами ВВК. Другими словами, процедура установления межмодульных (межкомпонентных) связей реализуется через сопоставление наблюдаемых характеристик текущего фрагмента с аналогичным (или сопоставимым) набором признаков других компонентов, коммутированных с данным модулем. Таким образом, фрагменты обучающей выборки выстраиваются в логическую последовательность или несколько последовательностей, отражающих структуру вычислительного комплекса в целом. Причем эта взаимосвязь устанавливается для каждого следующего временного отрезка, что позволяет выделить закономерности перехода от одного состояния к другому. Количество объединяемых фрагментов не ограничено.

Исходя из положения о гетерогенности вычислительного комплекса, разрабатывается схема прогнозирования, согласно которой после прогнозирования сбоев и отказов ФБ подключается отдельная искусственная нейронная сеть, предназначенная для синтеза результатов предыдущего шага. Ее основная задача заключается в выработке прогнозного значения относительно факта отказа целевого модуля (компонента) вычислительной системы. Такой подход обеспечивает значительную детализацию и глубину анализа, учитывая индивидуальные особенности каждого компонента системы и сводя к минимуму искажающее влияние гетерогенности структуры вычислительного комплекса на итоговое заключение о состоянии системы в целом.

Предлагаемый параллельный метод ассемблирования нейронных сетей предполагает построение нейросетевого каскада на основе выделенных фрагментов обучающих выборок. Подобная многослойная конструкция формирует иерархию уровней (ярусов каскада), число которых варьируется от двух и более в зависимости от сложности поставленной задачи и требуемой глубины анализа [20].

Проведение экспериментов. Для проверки корректности и эффективности предложенного параллельного метода каскадирования нейросетевых модулей был синтезирован экспериментальный комплекс: аппаратная часть – на базе высокопроизводительной кластерной вычислительной системы HP ENIGMA X000, установленной в ФГБОУ ВО «Вятский государственный университет», программная часть – прототип специализированной системы прогнозирования в среде MatLab. Система HP ENIGMA X000 содержит 288 узлов (node), большинство из которых реализованы на тонких серверах HP ProLiant 460 c. Техническое обслуживание и диагностика кластера включает: непрерывный контроль текущего состояния системы, систематический сбор и последующий детальный анализ служебных журналов событий (лог-файлов), а также выполнение специализированных тестовых процедур для идентификации потенциально опасных отклонений в работе подсистем.

Экспериментальный комплекс собран из четырех блэйдов и моделирует работу дуплексной вычислительной системы (рис. 4).

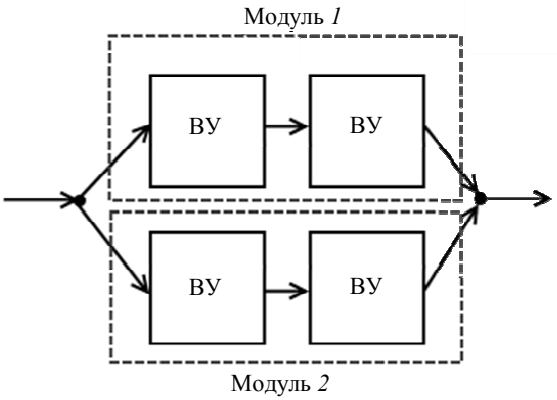

Рис. 3. Модель дуплексной вычислительной системы
Fig. 3. Duplex computing system model

Каждый вычислительный узел (ВУ) соответствует одной элементарной машине или одному функциональному блоку. Имеется два модуля, работающих в дуплексном режиме, причем в каждом модуле имеется конвейер из двух ВУ. Такая структура, при всей своей простоте, максимально информативна для исследования коммутационных зависимостей между функциональными блоками. Полноценный отказ всей системы возможен лишь при одновременном выходе из строя обоих вышеуказанных модулей. При этом остановка работы любого из конвейеров наступает вследствие единичного отказа хотя бы одного вычислительного узла, входящего в его состав.

Для воспроизведения реальной рабочей обстановки и активизации механизма возникновения отказов в рамках проводимого исследования система подверглась специальной нагрузочной

проверке, заключающейся в намеренном многократном запуске чрезмерного числа заявок сверх штатной пропускной способности, что позволило инициировать контролируемую последовательность событий, ведущую к выходу системы из рабочего состояния. При проведении эксперимента анализировались данные, снимаемые для каждого ВУ подсистемой мониторинга, входящей в управляющий кластерный пакет Oscar 4.0 (Enabled).

На начальной стадии эксперимента были сформированы обучающие выборки для отдельных блоков и модулей, входящих в состав комплекса. Базой для формирования выборок являются данные из «Служебного журнала событий» администратора кластера ВятГУ за период с мая 2024 по май 2025 г. При этом работа прогностической модели сводится к организации именно краткосрочных прогнозов (3–5 с). В пределах этого времени результат прогнозов поступает принимающему решения лицу, которое, в свою очередь, определяет возможность отключения блэйда.

Анализ результатов. Первый прототип специализированной системы прогнозирования был реализован на базе модели с архитектурой GRNN (General Regression Neural Network). В качестве изменяемого параметра, у обобщенно-регрессионной сети выбрана целевая ошибка. Результаты прогнозирования, полученные с использованием первого экспериментального прототипа приведены в табл. 3. В столбце «Прогноз GRNN», указана пара – (вероятность сбоя; вероятность отказа), причем число, близкое к 1, соответствует отсутствию сбоя/отказа, а число, близкое к 2, – появлению сбоя/отказа в работе вычислительного модуля. Целевая ошибка представляет собой заранее предустановленное расхождение между выходным значением модели и истинным значением метки для каждого конкретного примера.

Табл. 3. Результаты экспериментов (GRNN-сеть)
Tab. 3. Experimental results (GRNN network)

Целевая ошибка	Прогноз_GRNN	Факт (журнал событий)
1	(1.118; 1.045)	Норма
0.1	(1.043; 1.332)	Норма
0.2	(1.238; 1.332)	Сбой
0.3	(1.497; 1.560)	Сбой
…	…	…
0.01	(1.102; 1.467)	Отказ
0.02	(0.994; 1.506)	Норма
…	…	…
0.001	(0.998; 1.511)	Норма
0.002	(1.211; 1.332)	Норма
…	…	…
0.0001	(1.932; 1.878)	Сбой

Точность предсказания сбоев составила 72 %, а появления ситуации «Отказ» – 62 % случаев.

Второй вариант специализированной системы прогнозирования содержал единственный нейросетевой модуль с архитектурой LVQ (Learning Vector Quantization). Результаты прогнозирования, полученные с использованием второго экспериментального прототипа приведены в табл. 4. В качестве изменяемого параметра у нейронной сети векторного квантования выбрано количество нейронов в конкурентном слое. В столбце «Прогноз LVQ», указана прогнозируемая тройка – (Нормальная работа, Сбой, Отказ).

Табл. 4. Результаты экспериментов (LVQ-сеть)
Tab. 4. Experimental results (LVQ network)

Количество нейронов	Прогноз_LVQ	Факт (журнал событий)
1	(1, 0, 0)	Норма
5	(1, 0, 0)	*Отказ*
10	(1, 0, 0)	Норма
20	(1, 0, 0)	*Сбой*
30	(1, 0, 0)	Норма
40	(0, 1, 0)	*Отказ*
…	…	…
100	(0, 1, 1)	Отказ
200	(1, 0, 0)	Норма
300	(1, 0, 0)	Норма
400	(1, 0, 0)	*Сбой*
500	(0, 1, 0)	Сбой
…	…	…
1000	(1, 0, 0)	Норма

Точность предсказания сбоев составила 68 %, а появление ситуации «Отказ» угадано верно в 71 % случаев. Появление в прогнозе одновременно ситуаций «Сбой» и «Отказ» (строка 8) обусловлено тем, что на начальном этапе выхода из строя дуплексного вычислительного комплекса возможно появление сбоя в работе какого-либо функционального блока, затем перемежающегося отказа (многократных сбоев в работе нескольких ФБ) и в конце этого процесса – полный, окончательный отказ комплекса.

Для проверки корректности и эффективности предлагаемого авторами метода обучающая выборка была разделена на три фрагмента. Первые две выборки формируются на основе зарегистрированных технических параметров, характеризующих функционирование двух модулей вычислительного комплекса, резервирующих друг друга в дуплексном режиме. Третий фрагмент выборки синтезирован из вероятностных оценок отказа обеих подсистем одновременно, рассчитанных для одних и тех же автоматных временных интер-

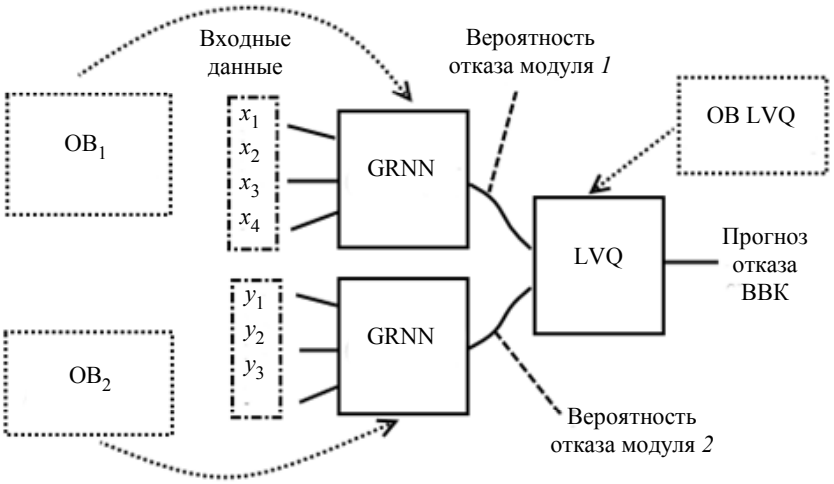

Рис. 5. Каскад нейросетевых модулей
Fig. 5. Cascade of neural network modules

валов. Реализация прогнозов осуществлялась посредством синтеза гетерогенного каскада нейросетевых модулей, организованно скоммутированных в два последовательных яруса (рис. 5).

Для прогнозирования сбоев и отказов модулей *1* и *2* указанного вычислительного комплекса в состав специализированной системы прогнозирования введены нейросетевые модули на основе обобщенной регрессионной сети (GRNN). А для прогнозирования работоспособности всей дуплексной системы в целом добавлен модуль на основе нейронной сети векторного квантования (LVQ). Результаты экспериментов с применением каскадного ансамбля приведены в табл. 5.

Табл. 5. Результаты экспериментов (каскад)
Tab. 5. Experimental results (cascade)

Количество нейронов второго яруса	Прогноз (каскад)	Факт (журнал событий)
1	(1, 0, 0)	Норма
5	(**1**, 0, 0)	***Сбой***
10	(1, 0, 0)	Норма
20	(1, 0, 0)	Норма
30	(0, 0, **1**)	**Отказ**
40	(1, 0, 0)	Норма
50	(1, 0, 0)	Норма
60	(1, 0, 0)	Норма
100	(1, 0, 0)	Норма
150	(1, 0, 0)	Норма
200	(1, 0, 0)	Норма
250	(1, 0, 0)	Норма
300	(0, 0, **1**)	**Отказ**
350	(1, 0, 0)	Норма
500	(0, **1**, 0)	***Сбой***
600	(1, 0, 0)	Норма
800	(1, 0, 0)	Норма
1000	(0, **1**, 0)	***Сбой***

В большинстве случаев (почти в 89 %) кратковременные сбои отдельных узлов экспериментального кластера были спрогнозированы верно. Вероятность отказов была спрогнозирована со 100 %-ной точностью. Конечно, необходимо учитывать, что на такой высокий показатель повлиял крайне незначительный срок прогнозирования – 1 неделя. Но такой же срок был установлен и для одномодульных прототипов системы.

По результатам экспериментов было разработано программное обеспечение с выводом результатов прогнозирования в специальную форму. На основе сообщений системы о возможных отказах отдельных «проблемных» функциональных блоков, такой сервер может быть отключен (Down) несмотря на то, что динамический мониторинг в соответствии с регламентом эксплуатации ВВК показал успешное прохождение всех тестов, запускаемых кластерным пакетом Oscar 4.0. Прогноз о полном отказе блэйда сформирован на базе появления многократных сбоев элементов и узлов в данном сервере. Экспертные оценки показывают, что многократные перемежающиеся сбои практически всегда приводят к окончательным устойчивым отказам аппаратной части. Своевременное отключение блэйда позволяет предупредить выход из строя дорогостоящего оборудования и заранее перераспределить вычислительные потоки на другие серверы, предотвратив потерю актуальных данных решаемой задачи.

Заключение. Использование искусственных нейронных сетей для прогнозирования нормальной работоспособности высокопроизводительных вычислительных комплексов в процессе их функционирования многообещающе. Но иногда результаты могут быть недостаточно точными или

даже ошибочными из-за отсутствия учета взаимовлияния отдельных факторов при эксплуатации масштабных вычислительных систем со сложной коммутацией. Предлагаемый в данной статье метод позволяет решить эти проблемы, выделяя взаимосвязи между элементами фрагментированных обучающих выборок и передавая промежуточные прогнозные значения между ярусами каскада. Фрагментация выборок приводит к уменьшению количества обрабатываемых данных на первом ярусе, что может значительно сократить аппаратные затраты и время на обучение нейросетей входного яруса. Кроме того, длительность обучения можно дополнительно уменьшить, используя разработанный авторами метод параллельного обучения всех модулей системы одновременно [21]. Более точный прогноз получается благодаря выделению специализированных нейросетевых модулей под каждый этап выполнения прогноза и благодаря тому, что данные для следующих слоев каскада формируются динамически и содержат уже обработанную достоверную информацию.

В высокопроизводительных вычислительных системах выделение фрагментов в обучающих и тестовых выборках для отдельных аппаратных компонентов с учетом взаимосвязи и взаимозависимостей между параметрами выборок позволяет применить метод агрегации сетей и построения каскада из нейросетевых модулей особенно эффективно. Проведение масштабного тестирования различных конфигураций нейросетевых каскадов с использованием обновленных вариантов обучающих выборок, предварительно подвергнутых процедуре разбиения на отдельные фрагменты, обеспечивает всестороннюю проверку надежности предлагаемого решения. Важно отметить, что обучающие выборки для последующих ярусов могут включать дополнительные уникальные параметры, отсутствующие в исходном наборе данных, что существенно повышает точность и качество процесса прогнозирования. Полученные результаты позволяют сделать вывод о высокой потенциальной эффективности данного подхода в рамках дальнейшего совершенствования технологий анализа больших объемов данных и обеспечения надежности функционирования масштабируемых высокопроизводительных цифровых комплексов.

Список литературы

1. Application of multivariate time-series model for high performance computing (HPC) fault prediction / X. Pei, M. Yuan, G. Mao, Z. Pang // PLOS One. 2023. Vol. 18, no. 10. Art. e0281519. P. 1–18. doi: 10.1371/journal.pone.0281519 (дата обращения: 11.07.2025).

2. A study of job failure prediction at job submit-state and job start-state in high-performance computing system: using decision tree algorithms / A. Banjongkan, W. Pongsena, K. Kerdprasop, N. Kerdprasop // J. of Advances in Inform. Technol. 2021. Vol. 12, no. 2. P. 84–92. doi: 10.12720/jait.12.2.84-92.

3. Исследование и разработка методов прогнозирования сбойных ситуаций в суперкомпьютерах в системе Octotron / С. И. Соболев, А. С. Антонов, П. А. Швец, Д. А. Никитенко, К. С. Стефанов, В. В. Воеводин, С. А. Жуматий // Сб. тр. XIII междунар. науч. конф. «Параллельные вычислительные технологии» (ПаВТ'2019). Калининград: Изд. центр ЮУрГУ (Челябинск), 2019. С. 396–401.

4. Pin-pointing node failures in HPC Systems / A. Das, P. Hargrove, F. Mueller, E. Roman // Lawrence Berkeley National Lab. 2020. URL: https://sc16.supercomputing.org/sc-archive/tech_poster/poster_files/post254s2-file3.pdf (дата обращения: 12.07.2025).

5. Анализ надежности кластерных систем высокой отказоустойчивости с напрямую подключенными устройствами хранения данных / А. Л. Панин, А. В. Хижняк, Е. И. Михненок, А. Ю. Липлялин // Докл. БГУИР. 2017. № 7 (109). С. 83–87.

6. Андрюхин Е. В., Ридли М. К., Правиков Д. И. Прогнозирование сбоев и отказов в распределенных системах управления на основе моделей прогнозирования временных рядов // Вопр. кибербезопасности. 2019. № 3(31). С. 24–32. doi: 10.21681/2311-3456-2019-3-24-32.

7. Богатырев В. А., Богатырев С. В., Богатырев А. В. Граничная оценка надежности кластерных систем на основе декомпозиции марковской модели при ограниченном восстановлении узлов с накоплением отказов // Науч.-техн. вестн. информационных технол., механики и оптики. 2025. Т. 25, № 3. С. 574–583. doi: 10.17586/2226-1494-2025-25-3-574-583.

8. Zhao Ch., Xi Yu. Real-time fault detection and stability enhancement mechanism based on large models // Intern. J. of Emerging Technol. and Advanced Appl. 2025. Vol. 2, no. 2. P. 1–12. doi: 10.62677/IJETAA.2502132.

9. Anomaly detection using autoencoders in high performance computing systems / A. Borghesi, A. Bartolini, M. Lombardi, M. Milano, L. Benini // Proc. of the Thirty-First AAAI Conf. on Innovative Appl. of Artificial Intelligence (IAAI-19). Honolulu, Hawaii, USA: AAAI Press, 2019. Vol. 33, no. 01. P. 9428–9433. doi: 10.1609/aaai.v33i01.33019428.

10. A hybrid model for predictive maintenance integrating gradient boosting and long short-term memory networks / N. Padmashri, V. V. Karthikeyan, P. Sumathi, R. Swathiramya // Intern. J. for Research in Eng. Appl. & Management (IJREAM). 2024. Vol. 10, no. 3. P. 44–59. doi: 10.35291/2454-9150.2024.0355.

11. Anomaly detection and anticipation in high performance computing systems / A. Borghesi, M. Molan, M. Milano, A. Bartolini // IEEE Transactions on Parallel and Distributed Syst. 2022. Vol. 33, no. 4. P. 739–750. doi: 10.1109/TPDS.2021.3082802.

12. Рудометкин В. А. Мониторинг и поиск неисправностей в распределенных высоконагруженных системах // Кибернетика и программирование. 2020. № 2. С. 1–6. doi: 10.25136/2644-5522.2020.2.32996.

13. An explainable model for fault detection in HPC Systems / M. Molan, A. Borghesi, F. Beneventi, M. Guarrasi, A. Bartolini // High Performance Comp. Lecture Notes in Comp. Sci. (LNCS). 2021. Vol. 12761. P. 378–391. doi: 10.1007/978-3-030-90539-2_25.

14. Сай В. К., Щербаков М. В. Классификационный подход на основе комбинации глубоких нейронных сетей для прогнозирования отказов сложных многообъектных систем // Моделирование, оптимизация и информ. технол. 2020. № 8(2). С. 1–11. doi: 10.26102/2310-6018/2020.29.2.037.

15. A comprehensive survey of machine learning and deep learning approaches for anomaly detection in high-performance computing systems / C. Ki, R. Sivakumar, J. Mulerikkal, A Binu, M. Gupta, T. Jan // The J. of Super-Comp. 2025. Vol. 81, no. 8. Art. 1032. doi: 10.1007/s11227-025-07503-4.

16. Failure prediction using machine learning in a virtualised HPC system and application / B. Mohammed, I. Awan, H. Ugail, M. Younas // Cluster Comp. 2019. Vol. 22, no. 6. P. 471–485. doi: 10.1007/s10586-019-02917-1.

17. Antici F., Borghesi A., Kiziltan Z. Online Job Failure Prediction in an HPC System // Euro-Par 2023: Parallel Proc. Workshops. Lecture Notes in Comp. Sci. (LNCS). 2024. Vol. 14352. P. 167–179. doi: 10.1007/978-3-031-48803-0_35.

18. Asmawi T. N. T., Ismail A., Shen J. Cloud failure prediction based on traditional machine learning and deep learning // J. of Cloud Comp. 2022. Vol. 11. Art. 47. doi: 10.1186/s13677-022-00327-0.

19. Хорошевский В. Г. Архитектура вычислительных систем // М.: Изд. МГТУ им. Н. Э. Бауман, 2008. 520 с.

20. Крутиков А. К., Мельцов В. Ю. Метод формирования многоярусной нейросетевой системы прогнозирования с возможностью реконфигурации // Изв. Юго-Западного гос. ун-та. 2024. Т. 28, № 4. С. 104–123. doi: 10.21869/2223-1560-2024-28-4-104-123.

21. Krutikov A. K., Meltsov V. Y., Strabykin D. A. Evaluation the efficienty of forecasting sports events using a cascade of artificial neural networks based on FPGA // Proc. of ElConRus-2022. SPb., RF: IEEE, 2022. P. 355–360. doi: 10.1109/ElConRus54750.2022.9755840.

Информация об авторах

Мельцов Василий Юрьевич – канд. техн. наук, доцент кафедры ЭВМ ВятГУ, ул. Московская, 36, г. Киров, 610000, Россия.
E-mail: meltsov69@mail.ru
https://orcid.org/0000-0001-5479-9979

Крутиков Александр Константинович – аспирант, ст. преподаватель кафедры ЭВМ ВятГУ, ул. Московская, 36, г. Киров, 610000, Россия.
E-mail: usr09603@vyatsu.ru
https://orcid.org/0000-0003-4142-7329

References

1. Application of multivariate time-series model for high performance computing (HPC) fault prediction / X. Pei, M. Yuan, G. Mao, Z. Pang // PLOS One. 2023. Vol. 18, no. 10. Art. e0281519. P. 1–18. doi: 10.1371/journal.pone.0281519 (data obrashhenija: 11.07.2025).

2. A study of job failure prediction at job submit-state and job start-state in high-performance computing system: using decision tree algorithms / A. Banjongkan, W. Pongsena, K. Kerdprasop, N. Kerdprasop // J. of Advances in Information Technology. 2021. Vol. 12, no. 2. P. 84–92. doi: 10.12720/jait.12.2.84-92.

3. Issledovanie i razrabotka metodov prognozirovanija sbojnyh situacij v superkomp'juterah v sisteme Octotron / S. I. Sobolev, A. S. Antonov, P. A. Shvec, D. A. Nikitenko, K. S. Stefanov, V. V. Voevodin, S. A. Zhumatij // Sb. tr. XIII mezhdunar. nauch. konf. «Parallel'nye vychislitel'nye tehnologii» (PaVT'2019). Kaliningrad: Izd. centr JuUrGU (Cheljabinsk), 2019. S. 396–401. (In Russ.).

4. Pin-pointing node failures in HPC Systems / A. Das, P. Hargrove, F. Mueller, E. Roman // Lawrence Berkeley National Lab. 2020. URL: https://sc16.supercomputing.org/sc-archive/tech_poster/poster_files/post254s2-file3.pdf (data obrashhenija: 12.07.2025).

5. Analiz nadezhnosti klasternyh sistem vysokoj otkazoustojchivosti s naprjamuju podkljuchennymi ustrojstvami hranenija dannyh / A. L. Panin, A. V. Hizhnjak, E. I. Mihnenok, A. Ju. Lipljalin // Dokl. BGUIR. 2017. № 7 (109). S. 83–87. (In Russ.).

6. Andrjuhin E. V., Ridli M. K., Pravikov D. I. Prognozirovanie sboev i otkazov v raspredelennyh sistemah upravlenija na osnove modelej prognozirovanija vremennyh rjadov // Vopr. kiberbezopasnosti. 2019. № 3(31). S. 24–32. doi: 10.21681/2311-3456-2019-3-24-32. (In Russ.).

7. Bogatyrev V. A., Bogatyrev S. V., Bogatyrev A. V. Granichnaja ocenka nadezhnosti klasternyh sistem na osnove dekompozicii markovskoj modeli pri ogranichennom vosstanovlenii uzlov s nakopleniem otkazov //

Nauch.-tehn. vestn. informacionnyh tehnologij, mehaniki i optiki. 2025. T. 25, № 3. S. 574–583. doi: 10.17586/2226-1494-2025-25-3-574-583. (In Russ.).

8. Zhao Ch., Xi Yu. Real-time fault detection and stability enhancement mechanism based on large models // Intern. J. of Emerging Technol. and Advanced Appl. 2025. Vol. 2, no. 2. P. 1–12. doi: 10.62677/IJETAA.2502132.

9. Anomaly detection using autoencoders in high performance computing systems / A. Borghesi, A. Bartolini, M. Lombardi, M. Milano, L. Benini // Proc. of the Thirty-First AAAI Conf. on Innovative Appl. of Artificial Intelligence (IAAI-19). Honolulu, Hawaii, USA: AAAI Press, 2019. Vol. 33, no. 01. P. 9428–9433. doi: 10.1609/aaai.v33i01.33019428.

10. A hybrid model for predictive maintenance integrating gradient boosting and long short-term memory networks / N. Padmashri, V. V. Karthikeyan, P. Sumathi, R. Swathiramya // Intern. J. for Research in Eng. Appl. & Management (IJREAM). 2024. Vol. 10, no. 3. P. 44–59. doi: 10.35291/2454-9150.2024.0355.

11. Anomaly detection and anticipation in high performance computing systems / A. Borghesi, M. Molan, M. Milano, A. Bartolini // IEEE Transactions on Parallel and Distributed Systems. 2022. Vol. 33, no. 4. P. 739–750. doi: 10.1109/TPDS.2021.3082802.

12. Rudometkin V. A. Monitoring i poisk neispravnostej v raspredeljonnyh vysokonagruzhennyh sistemah // Kibernetika i programmirovanie. 2020. № 2. S. 1–6. doi: 10.25136/2644-5522.2020.2.32996. (In Russ.).

13. An explainable model for fault detection in HPC Systems / M. Molan, A. Borghesi, F. Beneventi, M. Guarrasi, A. Bartolini // High Performance Comp. Lecture Notes in Comp. Sci. (LNCS). 2021. Vol. 12761. P. 378–391. doi: 10.1007/978-3-030-90539-2_25.

14. Saj V. K., Shherbakov M. V. Klassifikacionnyj podhod na osnove kombinacii glubokih nejronnyh setej dlja prognozirovanija otkazov slozhnyh mnogoob#ektnyh sistem // Modelirovanie, optimizacija i informacionnye tehnologii. 2020. № 8(2). S. 1–11. doi: 10.26102/2310-6018/2020.29.2.037. (In Russ.).

15. A comprehensive survey of machine learning and deep learning approaches for anomaly detection in high-performance computing systems / C. Ki, R. Sivakumar, J. Mulerikkal, A Binu, M. Gupta, T. Jan // The J. of Super-Computing. 2025. Vol. 81, no. 8. Art. 1032. doi: 10.1007/s11227-025-07503-4.

16. Failure prediction using machine learning in a virtualised HPC system and application / B. Mohammed, I. Awan, H. Ugail, M. Younas // Cluster Comp. 2019. Vol. 22, no. 6. P. 471–485. doi: 10.1007/s10586-019-02917-1.

17. Antici F., Borghesi A., Kiziltan Z. Online Job Failure Prediction in an HPC System // Euro-Par 2023: Parallel Proc. Workshops. Lecture Notes in Comp. Sci. (LNCS). 2024. Vol. 14352. P. 167–179. doi: 10.1007/978-3-031-48803-0_35.

18. Asmawi T. N. T., Ismail A., Shen J. Cloud failure prediction based on traditional machine learning and deep learning // J. of Cloud Comp. 2022. Vol. 11. Art. 47. doi: 10.1186/s13677-022-00327-0.

19. Horoshevskij V. G. Arhitektura vychislitel'nyh sistem // M.: Izd. MGTU im. N. Je. Bauman, 2008. 520 s. (In Russ.).

20. Krutikov A. K., Mel'cov V. Ju. Metod formirovanija mnogojarusnoj nejrosetevoj sistemy prognozirovanija s vozmozhnost'ju rekonfiguracii // Izv. Jugo-Zapadnogo gos. un-ta. 2024. T. 28, № 4. S. 104–123. doi: 10.21869/2223-1560-2024-28-4-104-123. (In Russ.).

21. Krutikov A. K., Meltsov V. Y., Strabykin D. A. Evaluation the efficienty of forecasting sports events using a cascade of artificial neural networks based on FPGA // Proc. of ElConRus-2022. SPb., RF: IEEE, 2022. P. 355–360. doi: 10.1109/ElConRus54750.2022.9755840.

Information about the authors

Vasily Yu. Meltsov – Cand. Sci. (Eng.), Associate Professor, Department of Computer Science, Vyatka State University, 36 Moskovskaya St., Kirov, 610000, Russia.
E-mail: meltsov69@mail.ru
https://orcid.org/0000-0001-5479-9979

Alexander K. Krutikov – postgraduate student, senior lecturer at the Department of Computer Science, Vyatka State University, 36 Moskovskaya St., Kirov, 610000, Russia.
E-mail: usr09603@vyatsu.ru
https://orcid.org/0000-0003-4142-7329

Статья поступила в редакцию 12.08.2025; принята к публикации после рецензирования 28.08.2025; опубликована онлайн 28.11.2025.

Submitted 12.08.2025; accepted 28.08.2025; published online 28.11.2025.