DOI https://doi.org/10.46845/2541-8254-2025-3(50)-42-42
УДК 004.415.2

ПРЕДСКАЗЫВАНИЕ ПРОИЗВОДИТЕЛЬНОСТИ КЛАСТЕРОВ

А. В. Смирнов, студент магистратуры 2-го курса
E-mail: aleksandr.smirnov@digital-klgtu.ru
ФГБОУ ВО «Калининградский государственный технический университет»

И. А. Шикота, студент магистратуры 2-го курса
E-mail: ilya.shickota@yandex.ru
ФГБОУ ВО «Калининградский государственный технический университет»

В. Д. Штерцер, студент магистратуры 2-го курса
E-mail: vadsterz2.0@gmail.com
ФГБОУ ВО «Калининградский государственный технический университет»,

А. В. Снытников, д.т.н., научный руководитель
E-mail: aleksej.snytnikov@klgtu.ru
ФГБОУ ВО «Калининградский государственный технический университет»

Оценка производительности кластеров – важная задача, так как она позволяет определить, какой сложности задачу и за какое время суперкомпьютер может решить. Однако существующие методы тестирования производительности не отображают фактическую производительность СуперЭВМ. В данной статье проведено исследование эффективности прогнозирования производительности суперкомпьютеров с использованием нейронных сетей.

Ключевые слова: *СуперЭВМ, суперкомпьютер, машинное обучение, кластер, сверточная нейронная сеть.*

ВВЕДЕНИЕ

Кластеры представляют собой мощные вычислительные системы, объединяющие ресурсы нескольких компьютеров для выполнения общих задач. Их архитектура и принципы работы позволяют достигать высокой производительности, отказоустойчивости и масштабируемости.

HPC – это технология, использующая кластеры мощных процессоров, которые работают параллельно для обработки массивных, многомерных наборов данных и решения сложных задач на чрезвычайно высокой скорости.

HPC решает некоторые из самых сложных современных вычислительных задач в режиме реального времени. Системы высокопроизводительных вычислений обычно работают со скоростью, более чем в миллион раз превышающей скорость самых быстрых настольных компьютеров, ноутбуков или серверов.

Суперкомпьютеры – специально созданные компьютеры с миллионами процессоров или процессорных ядер – уже несколько десятилетий играют важную роль в высокопроизводительных вычислениях. В отличие от мэйнфреймов, суперкомпьютеры гораздо быстрее и могут выполнять миллиарды операций с плавающей запятой за одну секунду.

Для того чтобы понимать, за какое время и какого рода задачи может решать суперкомпьютер, необходимо измерять его производительность. Но используемые для этого наборы тестов не показывают точное значение производительности.

ОБЪЕКТ ИССЛЕДОВАНИЯ

Объектом данного исследования являются *нейронные сети*.

ЦЕЛЬ И ЗАДАЧИ ИССЛЕДОВАНИЯ

Целью данного исследования является исследование эффективности применения моделей нейронной сети для прогнозирования производительности СуперЭВМ.

МЕТОДЫ ИССЛЕДОВАНИЯ

Суперкомпьютер – это тип компьютера с более высоким уровнем производительности по сравнению с компьютером общего назначения. Производительность суперкомпьютера обычно измеряется в операциях с плавающей точкой в секунду (FLOPS), а не в миллионах инструкций в секунду (MIPS).

По своей сути суперкомпьютер работает так же, как и обычный компьютер, но в огромных масштабах. Однако главное отличие заключается в том, что в суперкомпьютерах используются тысячи или даже миллионы процессоров, работающих параллельно, а не один процессор. Эти процессоры делят массивные вычислительные задачи на более мелкие части, решая их одновременно. Именно параллельная обработка позволяет суперкомпьютерам достигать потрясающих уровней производительности.

Типичный суперкомпьютер состоит из множества взаимосвязанных узлов, каждому из которых поручено выполнять часть общей нагрузки. Эта взаимосвязанная сеть, часто называемая узлами, позволяет машине эффективно распределять вычислительные задачи и обрабатывать их параллельно [1].

Оценка производительности суперкомпьютеров (СуперЭВМ) – это сложная задача, которая требует учета множества факторов.

Оценка производительности суперкомпьютеров – это комплексный процесс, направленный на измерение вычислительной мощности, эффективности работы памяти, коммуникационных возможностей и энергопотребления системы. Поскольку суперкомпьютеры используются для решения сложных научных, инженерных и промышленных задач, их тестирование требует специализированных подходов.

Основные цели оценки производительности

- Определение реальной вычислительной мощности:
  - измерение скорости выполнения операций (FLOPS);
  - проверка эффективности параллельных вычислений.
- Выявление узких мест:
  - анализ памяти, сети, ввода-вывода (I/O).
- Сравнение архитектур.
- Оптимизация энергопотребления.

Существуют различные методы и бенчмарки для оценки вычислительной мощности, эффективности и масштабируемости. Ниже приведены самые применяемые из них.

- Синтетические тесты (измерение пиковой производительности):
  - LINPACK (HPL)
    - решает систему линейных уравнений, используется в TOP500;
    - показывает Rmax (максимальная производительность) и Rpeak (теоретический максимум);
    - недостаток: не отражает реальные нагрузки, только «идеальные» условия;

    - HPCG (High-Performance Conjugate Gradient)
        - тестирует память и коммуникации через разреженные матрицы;
        - дает более реалистичную оценку, чем LINPACK.
- Прикладные бенчмарки (имитация реальных задач)
    - NAS Parallel Benchmarks (NPB)
        - моделирует задачи аэродинамики (CFD) через BT, SP, LU, FT, MG;
        - оценивает параллельную масштабируемость.
    - SPEC CPU/MPI
        - тестирует производительность CPU и MPI в реалистичных сценариях.
- Тесты памяти и коммуникаций
    - STREAM Benchmark
        - измеряет пропускную способность памяти (GB/s).
    - Intel MPI Benchmarks (IMB)
        - проверяет задержки и скорость обмена данными между узлами [2].

Так как целью данной работы является определение производительности суперкомпьютеров, за основу сравнения возьмем следующие тесты измерения производительности:

- LinPack [3];
- HPCG;
- NAS Parallel Benchmarks [4];
- SPEC.

Существующие решения во многом являются узконаправленными либо используют устаревшие технологии. Существующие бенчмарки либо определяют пиковую производительность, либо имеют узкую направленность (таблица 1) [5].

Таблица 1 – Существующие тесты производительности СуперЭВМ

Тест	Тип нагрузки	Что измеряет	Метрика	Плюсы	Минусы
HPLinPack	Линейная алгебра	Пиковая производительность (FLOPS)	Rmax, Rpeak (в FLOPS)	Простота	Не отражает реальные нагрузки
HPCG	Разреженные матрицы	Пропускная способность памяти, сеть	GFLOPS/TFLOPS	Близок к реальным задачам	Низкая абсолютная производительность
NAS NPB	Вычислительная гидродинамика	Параллельная эффективность	Время выполнения, скорость	Реалистичные научные задачи	Ограниченная область применения
SPEC CPU	Универсальные CPU-нагрузки	Производительность процессора	SPECrate, SPECspeed	Широкий охват задач	Не учитывает GPU/ускорители

Именно по этой причине авторами был разработан свой набор тестов для определения фактической производительности СуперЭВМ.

Однако это не уменьшает временные и вычислительные затраты, чего, в свою очередь, могла бы добиться обученная на данных суперкомпьютеров нейронная сеть.

На первом этапе исследования, результаты которого представлены в [6], был проведен сравнительный анализ существующих нейронных систем, в результате которого для дальнейшей работы была выбрана трапециевидная сверточная нейронная сеть.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Данные генерировались по следующему алгоритму:

- Подбирались следующие характеристики вычислительной системы:
  - общее количество ядер;
  - количество ядер на сокет;
  - частота процессора.
- Запускался комплекс тестирования с этими параметрами.
- В результате тестирования получали фактическую производительность при таких характеристиках ВС.
- Данные заносились в Excel-файл для дальнейшего анализа.
- После генерации определенного количества данных они использовались для обучения нейронной сети.
- Затем была проведена оптимизация параметров нейронной сети для получения максимальной точности предсказания.

Подобранные параметры нейронной сети под новый набор данных представлены ниже:

- Сетка(kernel size): 2
- Количество фильтров: [512, 128]
- Количество нейронов: [512, 256, 128, 64, 32]
- Функция оптимизации: Adam [7]
- Функция потери: Средняя абсолютная ошибка
- Функция активации: RELU [8]
- Количество эпох: 300
- batch size: 400

В таблице 2 представлены результаты показателей измерений прогнозирования производительности.

Таблица 2 – Результаты показателей измерений

Название	Коэффициент детерминации	Среднеквадратичная ошибка	Средняя абсолютная ошибка
Трапеционная сверточная нейронная сеть	0,85	0,001	0,007

В ходе работы было определено, что нейронные сети могут предсказывать фактическую производительность с высокой точностью и могут снизить расходы времени и вычислительных ресурсов.

ЗАКЛЮЧЕНИЕ

В ходе исследования был сделан обзор существующих тестов производительности СуперЭВМ, выделены их преимущества и недостатки, на основе которых был разработан комплекс тестирования, определяющий фактическую производительность, с помощью которого были сгенерированы данные для разработанной нейронной сети.

СПИСОК ЛИТЕРАТУРЫ

1. What is a Supercomputer? An Introduction to Super Computing [Электронный ресурс]. – URL: https://www.terakraft.no/post/what-is-a-supercomputer-a-comprehensive-guide (дата обращения: 08.05.2025).

2. Supercomputer Benchmarks. A comparison of HPL, HPCG, and HPGMG and their Utility for the TOP500 [Электронный ресурс]. – URL: http://fs.hlrs.de/projects/teraflop/24thWorkshop_talks/Erich_Strohmaier_Benchmarks_WSSP24.pdf (дата обращения: 08.05.2025).

3. The LINPACK Benchmark: An Explanation / J. J. Dongarra. – 1988.

4. The NAS Parallel Benchmarks / D. H. Bailey. – 1991.

5. High Performance Computing: Modern Systems and Practices / T. Sterling, M. Brodowicz, M. Anderson. – 2017.

6. Исследование эффективности применения глубокого обучения для прогнозирования производительности суперЭВМ / А. В. Смирнов, И. А. Шикота, В. Д. Штерцер, А. В. Снытников // Вестник молодежной науки. – 2024. – № 5(47). – С. 6. – DOI 10.46845/2541-8254-2024-5(47)-6-6. – EDN YOVIZI.

7. Adam: A method for stochastic optimization / D. P. Kingma, J. Lei Ba. – 2014.

8. ReLU (Rectified Linear Unit) Activation Function [Электронный ресурс]. – URL: https://iq.opengenus.org/relu-activation/ (дата обращения: 08.05.2025).

INVESTIGATING THE EFFECTIVENESS OF APPLYING DEEP LEARNING FOR PERFORMANCE PREDICTION

A. V. Smirnov, 2nd year master's student
E-mail: aleksandr.smirnov@digital-klgtu.ru
Kaliningrad State Technical University

I. A. Shikota, 2nd year master's student
E-mail: ilya.shickota@yandex.ru
Kaliningrad State Technical University

V. D. Shterzer, 2nd year master's student
E-mail: vadsterz2.0@gmail.com
Kaliningrad State Technical University

A. V. Snytnikov, Professor
E-mail: aleksej.snytnikov@klgtu.ru
Kaliningrad State Technical University

A cluster is a group of computers (nodes) united into a single system to perform common computing tasks. Clusters are used to improve the performance, fault tolerance and scalability of computing systems. A supercomputer is a typical example of clusters, used to perform complex computations at speeds significantly faster than general-purpose computers. Unlike conventional PCs or servers, supercomputers are optimised for high-performance computing (HPC) tasks such as scientific modelling, weather forecasting, quantum mechanics and artificial intelligence training.

Key words: *Supercomputer, supercomputer, machine learning, cluster, convolutional neural network.*