

Отчет по проекту: “Стохастический бинарный нейрон как платформа для вероятностных вычислений”

Введение

Развитие CMOS электроники, вызванное непрерывной миниатюризацией транзисторов и соответствующим экспоненциальным ростом вычислительных мощностей, на протяжении десятилетий было движущей силой технологического прогресса. Однако, когда масштабирование транзисторов достигло своих физических и экономических пределов, случился бум развития машинного обучения (ML) и искусственного интеллекта (AI). Возрастающий спрос на вычислительные ресурсы для обучения и поддержки крупных ML-моделей привёл к заметной нехватке энергоэффективных аппаратных компонентов. Помимо этого, растущая роль AI в таких отраслях, как автономный транспорт, подчёркивает глобальный характер этой проблемы, выходящий за рамки локальной инфраструктуры.

Инновации в полупроводниковой электронике продолжаются благодаря усовершенствованию традиционных транзисторных технологий, включая 3D-гетерогенную интеграцию и использование двумерных (2D) материалов для транзисторов и интерфейсов. Дальнейшие инновации в физике транзисторов, благодаря таким эффектам как отрицательная ёмкость, также представляют перспективы повышения эффективности. Альтернативный подход заключается в расширении экосистемы CMOS за счёт новых несиликоновых нанотехнологий. Этот подход позволяет разрабатывать гетерогенные архитектуры CMOS + X, где X — это совместимая с CMOS нанотехнология, например, магнитные, ферроэлектрические, мемристивные или фотонные системы. Одним из наиболее перспективных направлений, которым занимается наша лаборатория ProFM, является интеграция технологии CMOS с основанными на двумерных сверхпроводниках стохастическими бинарными нейронами (СБН).

Стохастические бинарные нейроны

Традиционные транзисторы, составляющие основу современной вычислительной техники, работают детерминированно: они переключаются между бинарными состояниями на основе чётко определённых электронных входных сигналов. С другой стороны, кубиты используют квантовую суперпозицию между состояниями 0 и 1, что позволяет реализовывать принципиально новые вычислительные парадигмы, но требует экстремальных условий окружающей среды.

СБН занимает промежуточное положение между этими двумя технологиями (рис. 1). В отличие от детерминированных транзисторов, СБН переключается между бинарными состояниями с определённой вероятностью, используя физический шум, а не жёсткие пороговые значения напряжения. Для создания СБН мы используем двумерные нанопровода из нитрида ниobia (NbN), являющегося сверхпроводящим материалом второго рода. Эти нанопровода демонстрируют резкий переход между сверхпроводящим и нормальным состояниями при превышении критического тока. Однако этот переход не является строго детерминированным — квантовые эффекты и тепловой шум вносят в него элемент случайности.

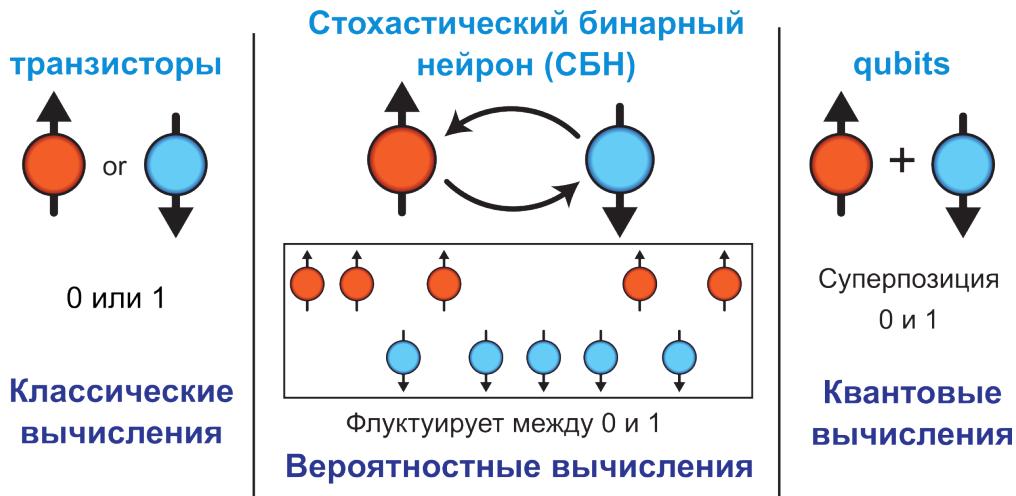


Рисунок 1: В каждом столбце - схематическое изображение базовых вычислительных компонентов: бит(слева), СБН (в середине) кубит (справа)

Эта природная стохастичность делает двумерные NbN нанопровода перспективной платформой для вероятностных вычислений, где случайность — это преимущество, а не недостаток.

Вероятностные вычисления и роль СБН

В основе вероятностных вычислений лежит СБН — аппаратный компонент, который использует внутренний физический шум для выполнения вероятностных вычислений. Исследования СБН начались на уровне устройств и физики, сосредотачиваясь на свойствах сверхпроводящих нанопроводов NbN. Эти материалы естественным образом проявляют стохастические переключения, что делает их подходящими для архитектур вероятностных вычислений. Ключевой принцип этих исследований заключается в прямом соответствии между физической стохастичностью и математическими моделями вероятностных алгоритмов, такими как методы Монте-Карло и машины Больцмана.

Методы Монте-Карло представляют собой класс вычислительных алгоритмов, основанных на случайных выборках для решения задач, которые, хотя и являются детерминированными в принципе, слишком сложны для точного аналитического решения. Эти методы широко применяются в физике, финансах и AI для оптимизации, интеграции и вероятностных выводов. Их сила заключается в приближённом поиске решений путём исследования возможных состояний через случайные процессы, что делает их идеальными для аппаратных реализаций с естественной стохастичностью.

Машины Больцмана, в свою очередь, представляют собой класс стохастических нейронных сетей, основанных на принципах минимизации энергии. Они вдохновлены статистической механикой и представляют собой сеть взаимосвязанных узлов, обновляющих свои состояния вероятностным образом в зависимости от взаимодействий. Возможность прямой аппаратной реализации вероятностного переключения позволяет СБН ускорять энергоэффективные модели, такие как машины Больцмана, делая их значительно более мощными по сравнению с традиционными реализациями на универсальных

процессорах.

Масштабирование вероятностных вычислений

Помимо разработки отдельных СБН-устройств, лаборатория ProFM активно исследует архитектурные инновации для масштабирования вероятностных вычислений. Исследуются такие методы, как сжатие и конвейерная обработка, для повышения эффективности и производительности. Подобно квантовым вычислениям, где прогресс зависит от междисциплинарного сотрудничества между аппаратными средствами, архитектурами и алгоритмами, вероятностные вычисления требуют аналогичного комплексного подхода.

Расширяя границы вычислений на основе СБН, лаборатория ProFM стремится к созданию высокопроизводительных, энергоэффективных вычислительных архитектур, которые могут органично интегрироваться с существующими полупроводниковыми технологиями.

Факторизация целых чисел как прототипная задача

Область адиабатических квантовых вычислений (АКВ) решает сложные задачи оптимизации, создавая сети кубитов, в которых взаимодействия между ними сконструированы таким образом, чтобы энергия системы E соответствовала функции стоимости рассматриваемой задачи. Один из алгоритмов представляет факторизацию целых чисел как задачу оптимизации, записывая множители X и Y в двоичном виде и определяя функцию стоимости:

$$E(x_P, \dots, x_1; y_Q, \dots, y_1) = \left[\left(\sum_{p=0}^P 2^p x_p \right) \left(\sum_{q=0}^Q 2^q y_q \right) - F \right]^2$$

где $x_0 = 1$, $y_0 = 1$, а P и Q обозначают количество бит, необходимых для представления X и Y , соответственно. Таким образом, состояние с наименьшей энергией соответствует конфигурации кубитов $\{x_p, \dots, x_1, y_q, \dots, y_1\}$, при которой произведение XY равно F .

В общем случае E включает в себя члены вида $x_p y_q x_r y_s$, требующие до четырёхчастичных взаимодействий. Этот алгоритм не требует когерентности, но при реализации в АКВ нуждается в дополнительных битах для представления многочастичных взаимодействий. В вероятностных вычислениях многочастичные взаимодействия реализуются электрически, исключая необходимость в дополнительных компонентах.

СБН выступают в качестве базовых элементов стохастических нейронных сетей. Поведение отдельных СБН определяется вероятностным механизмом: выходной сигнал m_i принимает значения 0 или 1 с вероятностями P_0 и P_1 соответственно. Эти вероятности зависят от нормированного входного сигнала I_i : при $I_i = 0$ нейрон с равной вероятностью выдаёт 0 или 1 ($P_0 = P_1 = 0.5$), при достаточно положительном I_i выход принудительно устанавливается в 1 ($P_0 = 0, P_1 = 1$), а при сильно отрицательном I_i — в 0

$(P_0 = 1, P_1 = 0)$. Этот стохастический переключательный механизм играет ключевую роль в реализации стохастических нейронных сетей для оптимизационных и обучающих задач.

$$m_i = \vartheta[\sigma(I_i) - r]$$

где ϑ — единичная ступенчатая функция, σ — сигмоидальная функция, r — случайное число, равномерно распределённое в диапазоне от 0 до 1, а вход I_i определяется синаптической функцией (описанной ниже). Таким образом, стохастические нейронные сети требуют естественного элемента, обладающего значительной нестабильностью, но при этом контролируемого.

Цель данной задачи — показать, что сложные задачи оптимизации могут быть решены в общем виде с использованием корреляции между несколькими естественно СБН.

В случае факторизации целых чисел мы используем функцию стоимости, представленную уравнением (1), для оценки входных функций. В качестве первого теста мы рассматриваем факторизацию 35 с использованием четырёх стохастических нейронных сетей ($P = 2, Q = 2$). В предлагаемом алгоритме синаптические входы включают нелинейные члены, которые эффективно обеспечивают как трёх-, так и четырёх-нейронные взаимодействия, наряду с традиционными линейными взаимодействиями между двумя нейронами. Соответственно, число до 2^{n+2} может быть закодировано согласно уравнению (1) с использованием n СБН. Этот подход требует меньше нейронов по сравнению с текущими схемами АКВ, главным образом благодаря дополнительной гибкости, обеспечиваемой нелинейными синапсами.

На рисунке 2 представлены трёхмерные гистограммы временных флюктуаций для пар чисел $\{x_2, x_1, 1\}$ и $\{y_2, y_1, 1\}$. Они изображены справа от некоррелированного состояния, полученного при установке всех входных функций в ноль. В некоррелированном состоянии (левая панель) СБН флюктуируют независимо. Однако при подаче ненулевого входного сигнала на сеть возникают два выраженных пика в точках $(5, 7)$ и $(7, 5)$, что демонстрирует корректную факторизацию 35 на 5 и 7 (правая панель).

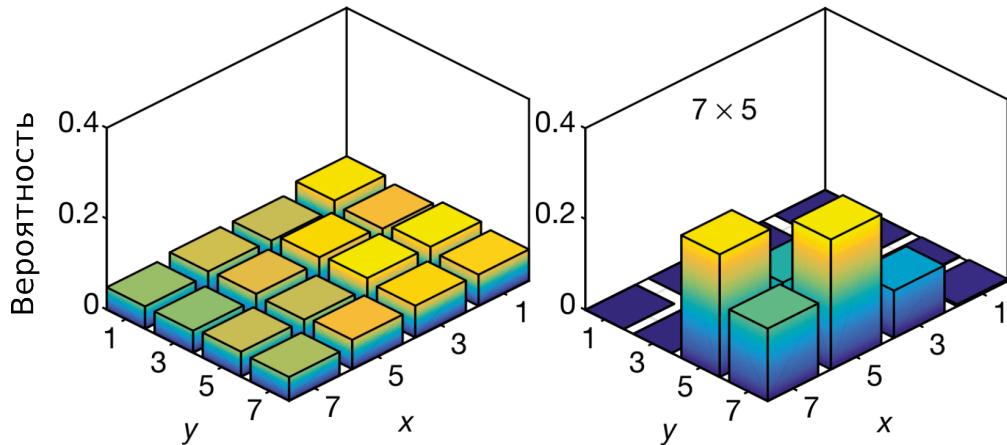


Рисунок 2: Некоррелированное (слева) и коррелированное (справа) состояние системы при использовании четырёх СБН для факторизации $35 = 5 \times 7 = 7 \times 5$ ($P = 2, Q = 2$).

Важно отметить, что детерминированные алгоритмы, реализуемые на полностью цифровых CMOS-системах, специализированы для выполнения факторизации. Однако по мере увеличения размера задачи таким системам требуется значительно больше времени для нахождения точного решения. С другой стороны, в случаях, когда допустимы приближённые решения, растёт интерес к аппаратным средствам, поддерживающим методы вероятностных вычислений.

Физические принципы

Фазовый переход в NbN

Стохастический характер перехода NbN в нормальное состояние обусловлен квантовыми и тепловыми флуктуациями, а также движением вихрей Абрикосова, которые могут временно закрепляться на дефектах и затем освобождаются. Иными словами, при приближении тока к критическому время жизни нанопровода в сверхпроводящем состоянии уменьшается из-за увеличения вероятности случайного перехода. После перехода в нормальное состояние материал нагревается за счет джоулева тепла, что предотвращает немедленный возврат в сверхпроводящее состояние при снижении тока ниже критического. Для восстановления сверхпроводимости ток должен быть существенно ниже критического. Это приводит к образованию гистерезисной петли на графике зависимости сопротивления от тока рис. 3.

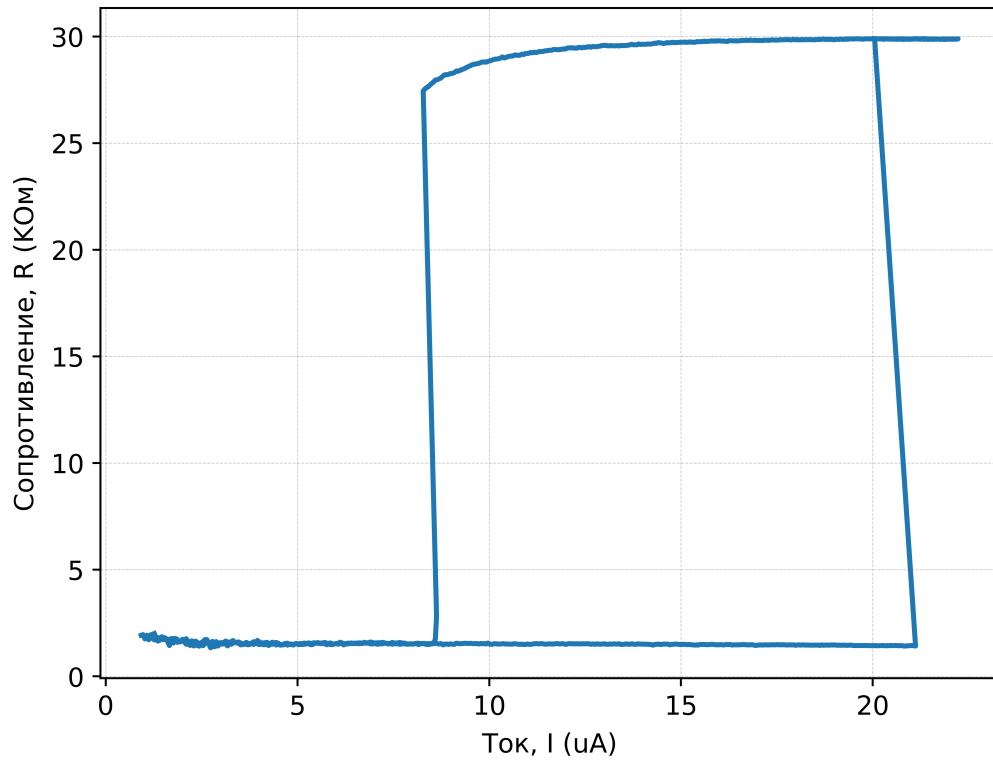


Рисунок 3: Гистерезис в зависимости сопротивления от тока, измеренный на нанопроводе NbN

Управление фазовым переходом в СБН

Критический ток в нанопроводе из NbN при температуре 4.5 К не превышает 50 мкА. При превышении критического тока сверхпроводимость разрушается, и сопротивление резко возрастает с близкого к нулю до нескольких десятков килоом. Сверхпроводящее и нормальное состояния нанопровода NbN могут быть интерпретированы как два состояния бинарного нейрона. Для вычислений на таком нейроне необходимо реализовать три ключевых условия:

1. Контроль вероятности перехода между состояниями.
2. Достоверная фиксация перехода в нормальное состояние.
3. Возможность многократного повторения процесса перехода и фиксации.

Для достижения этих условий на нанопровод подается синусоидальный ток с регулируемой амплитудой. При увеличении амплитуды до значений, близких к критическому току, вероятность перехода в нормальное состояние возрастает за счет стохастических процессов. Таким образом, реализуется первое требование. После перехода в нормальное состояние система остается в нем до тех пор, пока ток не снизится практически до нуля, что создает временное окно для фиксации перехода (второе требование). На следующем такте подачи синусоидального тока система возвращается в начальное состояние, что позволяет повторять вычисления многократно (третье требование).

Дополнительным преимуществом NbN является его технологичность, механическая прочность, а также высокая скорость переключения. Аналогичные материалы используются в однофотонных детекторах, где достигается быстродействие до 50 МГц.

Физическая реализация СБН

Для физической реализации стохастического бинарного нейрона (СБН) в нашей лаборатории был выбран сверхпроводящий нитрид ниобия (NbN). Структура представляет собой тонкую полоску шириной 150–300 нм и длиной 5–15 мкм с подведенными контактами с обеих сторон (рис. 4). Из-за малых размеров полоски тепло, выделяющееся при переходе в нормальное состояние, не успевает рассеяться мгновенно. В результате NbN находится в режиме стохастических переключений между сверхпроводящим и нормальным состояниями, причем вероятность переключения определяется амплитудой приложенного напряжения. Это свойство лежит в основе работы стохастического бинарного нейрона.

Процесс фабрикации

Фабрикация СБН производится на кремниевой подложке, покрытой слоем оксида кремния. Процесс включает несколько ключевых этапов:

1. Формирование контактных площадок.

На подложку напыляются золотые контакты и площадки, необходимые для подключения чипа к измерительному оборудованию (рис. 5).

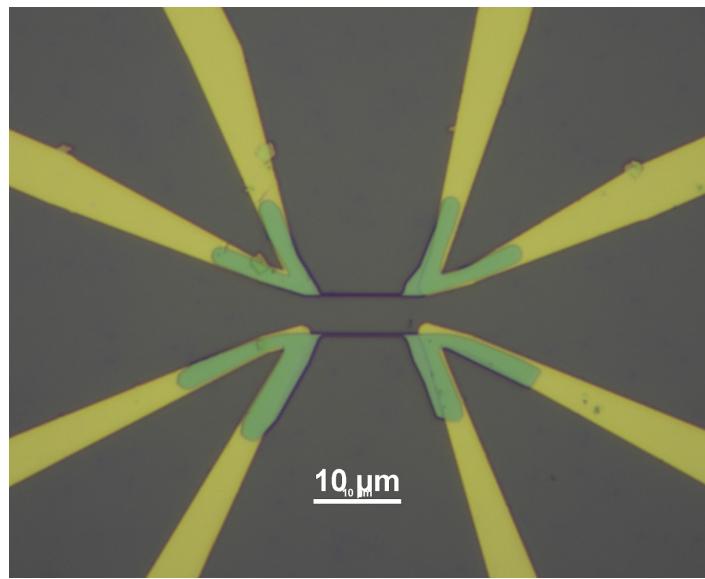


Рисунок 4: Оптическое изображение двух параллельных СБН из нитрида ниобия шириной 250 нм и 300 нм, с подведенными золотыми контактами. Показана центральная область чипа, размерная шкала 10 мкм. Тонкие полоски в середине и области зеленого цвета – нитрид ниобия, дорожки желтого цвета – золото.

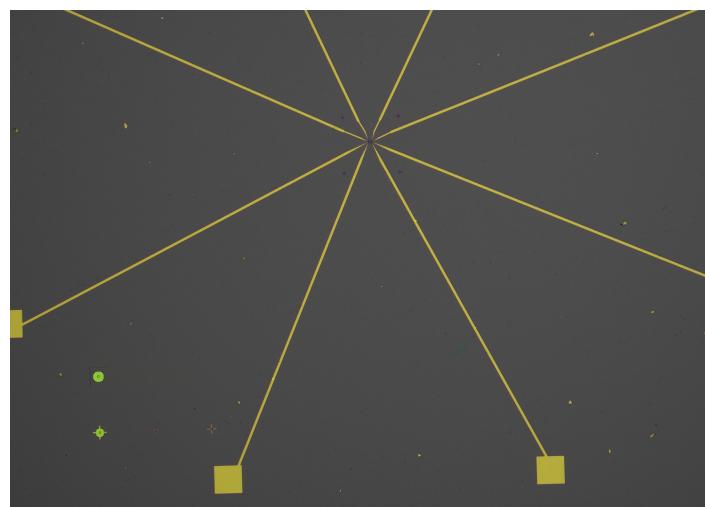


Рисунок 5: Оптическое изображение чипа с золотыми контактами и площадками. Размер области ~2×1,5 мм. На золотые площадки в дальнейшем устанавливаются золотые проволоки, идущие на переходную плату для подключения чипа к измерительному оборудованию.

2. Напыление нитрида ниобия.

Поверх контактов напыляется слой нитрида ниобия (20–25 нм) с подслоем из ниобия (5 нм). Оба слоя наносятся последовательно в рамках одного технологического процесса через предварительно подготовленную маску. Это позволяет сформировать узкие полоски NbN с перекрытием на золотых контактах (рис. 6).

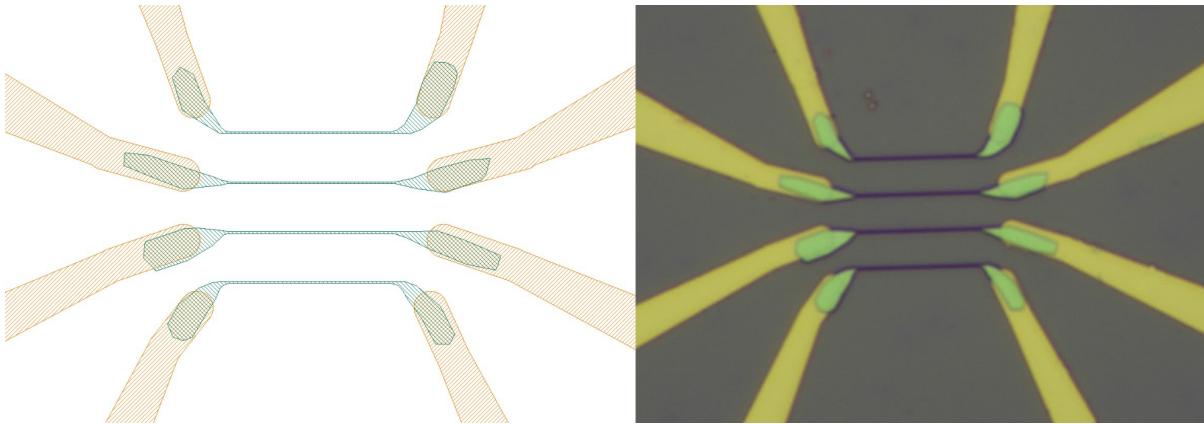


Рисунок 6: Разработанный дизайн (слева) и оптическое изображение готового чипа (справа) из четырех параллельно расположенных СБН шириной 250 нм с подведенными золотыми контактами. Показана центральная область чипа, примерный размер области на оптическом изображении 60x40 мкм. Тонкие полоски в середине, и области зеленого цвета - нитрид ниобия, дорожки желтого цвета - золото.

3. Оптимизация толщины нитрида ниобия.

В первых версиях устройства толщина NbN составляла ~50 нм, но измерения показали отсутствие гистерезиса – ключевого эффекта, необходимого для стохастических переключений. Уменьшение толщины NbN вдвое позволило добиться требуемого гистерезиса.

4. Электронно-лучевая литография.

Формирование масок для напыления золота и NbN выполняется методом электронно-лучевой литографии. Этот процесс подвержен уширению рисунка, зависящему от параметров экспозиции, что может приводить к расхождению реальной ширины полосок с заданной в дизайне. Были проведены тесты для точной калибровки размеров структур.

5. Финальная сборка.

На завершающем этапе на золотые площадки чипа устанавливаются золотые проволочные соединения, соединяющие сигнальные линии СБН с переходной платой и измерительным оборудованием.

Альтернативный метод изготовления

Для улучшения параметров нейрона исследуется альтернативный метод фабрикации, основанный на вытравливании полосок из цельного слоя NbN, нанесенного на всю кремниевую подложку. Такой подход

позволяет:

- Использовать более высокие температуры напыления ($>300^{\circ}\text{C}$), что улучшает кристаллическую структуру материала.
- Повысить однородность структуры, что потенциально увеличит стабильность работы нейрона.

Травление производится методом сухого плазмохимического травления в смеси газов Ar и SF₆. Мaska для травления создается методом электронно-лучевой литографии.

Оптимизация работы на высоких частотах

Для работы СБН на высоких частотах (с минимальными потерями мощности) между его выводами и золотыми площадками на чипе добавлены согласующие тэперы – дорожки специального дизайна с плавным расширением ширины. Они уменьшают отражение сигналов на высоких частотах, улучшая согласование импедансов. Пример таких тэперов показан на рис. 7.

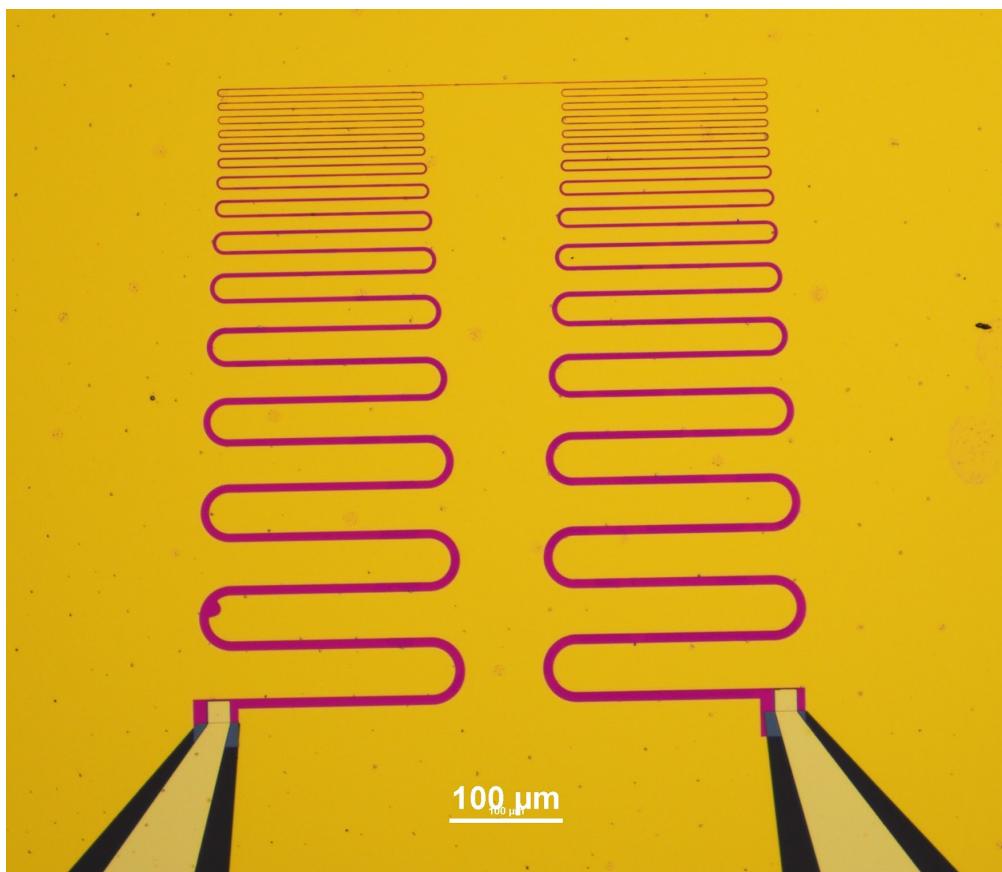


Рисунок 7: Оптическое изображение согласующих тэперов, размерная шкала 100 мкм.

Характеризация полученных СБН

Экспериментальная установка

Эксперимент проводился на нанопроводе NbN при частоте тактового генератора 500 кГц. Генератор позволял изменять амплитуду выходного сигнала и его постоянное смещение. Сигнал подавался через сверхпроводящий нанопровод на вход осциллографа с нагрузкой 50 Ом. При переходе нанопровода в нормальное состояние его сопротивление возрастало, что приводило к падению тока и фиксировалось на осциллографе в виде провала сигнала (рис. 8).

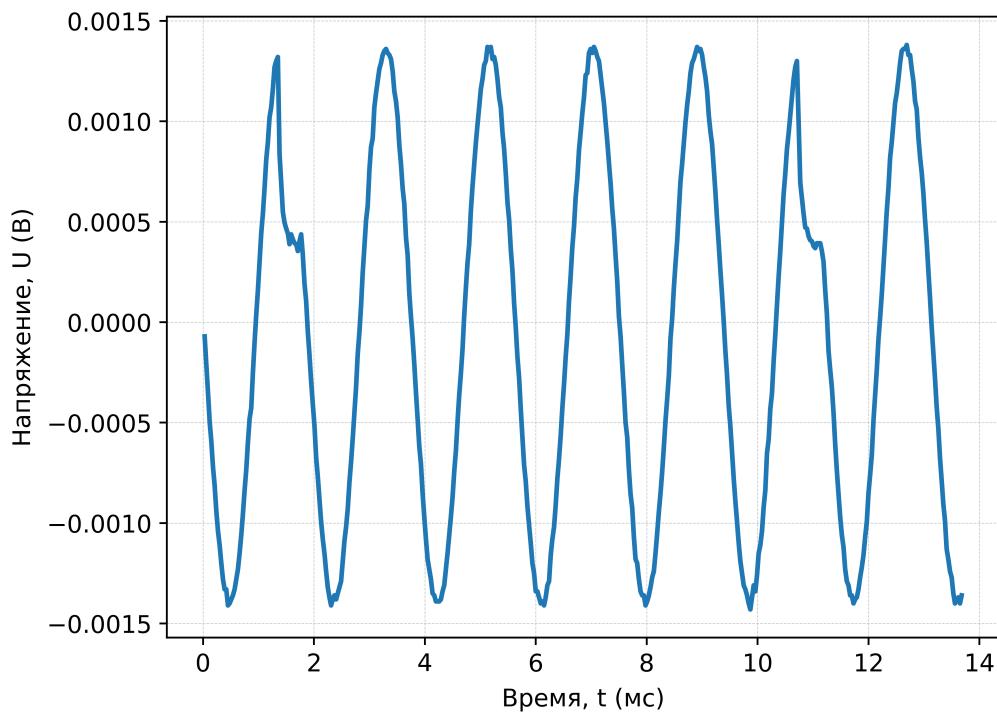


Рисунок 8: Сигнал с осциллографа, на котором зафиксировано два перехода в нормальное состояние

Была снята зависимость выходного сигнала от амплитуды синусоиды и смещения. Далее данные обрабатывались, и для каждого такта определялся факт перехода или его отсутствие. Полученные бинарные последовательности использовались для анализа вероятности перехода в зависимости от управляющего сигнала. Также была рассчитана автокорреляционная функция (рис. 9 и рис. 10) и проведены статистические тесты.

Оценка случайности переходов

Для оценки случайности бинарных последовательностей использовался пакет NIST “A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications”. Из 16 тестов успешно проходилось от 2 до 4, в зависимости от анализируемой серии данных. При этом некоторые тесты заведомо не могут быть пройдены, так как проверяют равновероятное распределение двух состояний, что не соответствует физике процесса.

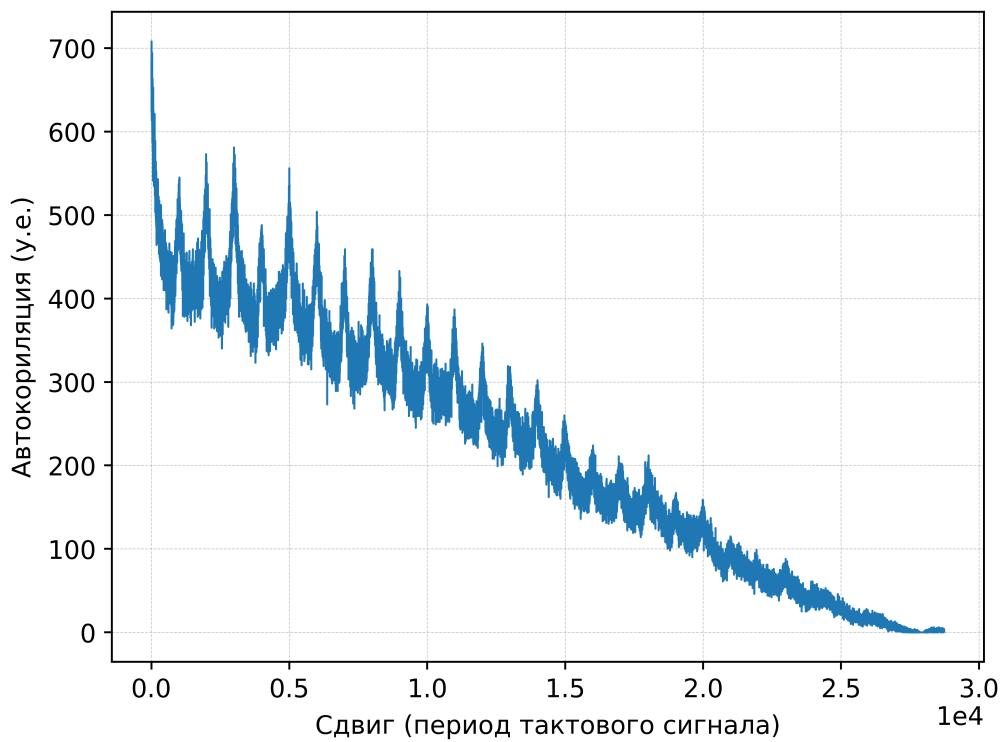


Рисунок 9: Пример автокорреляционной функции для данных с явной периодичностью.

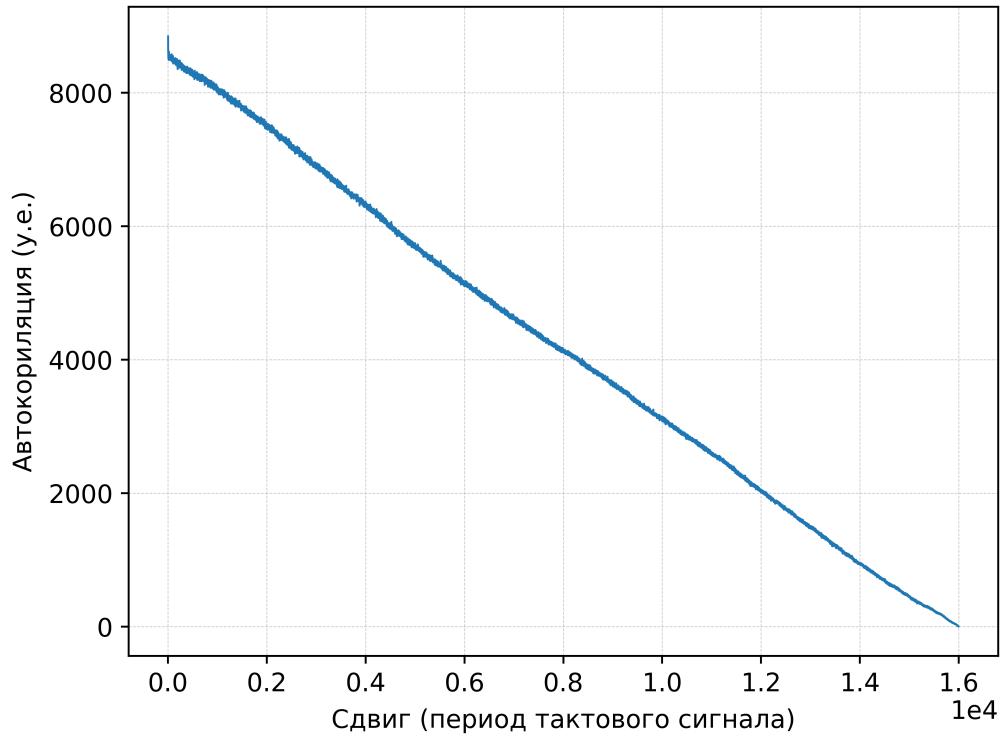


Рисунок 10: Пример автокорреляционной функции для данных близких к случайным

Малое количество успешно пройденных тестов, вероятно, связано с присутствием внешних детерминированных помех, влияющих на процесс перехода. В настоящее время проводится модернизация измерительного стенда с учетом возможных источников этих помех.

Заключение

Стохастические бинарные нейроны представляют собой перспективное направление в области вероятностных вычислений, объединяя физическую стохастичность сверхпроводниковых нанопроводов с эффективными алгоритмическими методами, такими как машины Больцмана и методы Монте-Карло. Разработка и исследование таких устройств позволяют не только глубже понять фундаментальные принципы фазовых переходов в низкоразмерных материалах, но и создать энергоэффективные вычислительные системы нового поколения.

Реализация СБН на основе нанопроводов NbN открывает возможности для аппаратного ускорения стохастических вычислений, а оптимизация их конструкции и технологии изготовления позволит улучшить стабильность работы и расширить диапазон применений. В дальнейшем масштабирование архитектур на основе СБН может привести к созданию специализированных процессоров, способных решать сложные вероятностные задачи быстрее и с меньшими затратами энергии, чем традиционные вычислительные системы.

Таким образом, работы, проводимые в лаборатории ProFM, закладывают основу для новых вычислительных парадигм, сочетающих преимущества квантовых и классических методов обработки информации.