NVIDIA Vera Rubin как новый стандарт инфраструктуры для Искусственного Интеллекта

Д.Е. Намиот, В.А. Сухомлин

Аннотация **- В статье анализируется новая платформа NVIDIA Vera Rubin, представленная в 2026 году и позиционируемая компанией как качественный скачок в построении вычислительной инфраструктуры для искусственного интеллекта. В отличие от традиционных подходов, ориентированных на отдельные чипы, платформа рассматривается как целостная система, объединяющая процессоры Vera (Arm), графические ускорители Rubin, высокоскоростные интерфейсы NVLink, сетевые адаптеры ConnectX-9 и программируемые DPU BlueField-4. Особое внимание уделяется аппаратной поддержке агентного ИИ, включая пространственную многопоточность, распределенное кэширование ключ-значение и масштабируемую стоечную архитектуру NVL72. Отдельный раздел посвящен применению цифровых двойников на базе Omniverse DSX для проектирования и эксплуатации крупных ИИ-фабрик. Авторы делают вывод, что платформа Vera Rubin знаменует переход от оценки производительности по пиковым FLOPS к системной оптимизации пропускной способности памяти и сети, задавая новые стандарты для инфраструктурных решений в области искусственного интеллекта.**

Ключевые слова- **NVIDIA, Искусственный интеллект, Vera Rubin.**

I. Введение

На CES 2026 компания NVIDIA представила новую компьютерную архитектуру Vera Rubin[1], хотя рынок ждал этот анонс ближе к концу года. Компания позиционирует платформу как следующий большой скачок и претендует с этим решением не просто на рост мощности, а на качественное изменение всей вычислительной инфраструктуры для ИИ.

Центральный элемент системы – это GPU Rubin[2]. NVIDIA утверждает, что этот процессор обеспечивает пятикратный прирост производительности в задачах обучения ИИ по сравнению с предыдущей системой (Blackwell). Однако ключевая идея платформы Vera Rubin не в одном чипе, а в полностью связанной системе, которая объединяет шесть компонент: CPU Vera, GPU Rubin, высокопроизводительную компьютерную шину NVLink, сетевой процессор DPU BlueField-4 и сетевой адаптер ConnectX-9. И все это проектировалось именно как единое целое, а не набор отдельных ускорителей.

По расчетам NVIDIA, переход на Vera Rubin позволит дата-центрам сократить количество GPU на 75%, сохранив или даже увеличив вычислительную мощность. Компания заявляет о десятикратном снижении совокупных затрат на вывод по сравнению с инфраструктурой на своей предыдущей платформе[3]. Речь идет об общей стоимости: «железо», энергопотребление, затраты на сеть и обслуживание.

Оставшаяся часть статьи структурирована следующим образом. В разделе II приведен обзор структуры платформы. В разделе III мы обсуждаем уникальность предложения NVIDIA. И раздел IV посвящен использованию эмуляторов при проектировании аппаратных платформ.

II. Структура платформы

Vera Rubin™ — это полноценная системная платформа с шестью ключевыми компонентами:

- NVIDIA Vera™ — Arm-процессор для ИИ-фабрик (88 ядер Olympus, Armv9.2);
- Rubin™ GPU — ускоритель нового поколения с HBM4 и NVLink 6;
- NVLink 6 Switch — коммутатор с жидкостным охлаждением;
- ConnectX-9 SuperNIC — сетевой адаптер с поддержкой Ethernet Photonics;
- BlueField-4 DPU — для обработки сетевых и хранилищных задач;
- Spectrum-6 Ethernet Switch — коммутатор для кластерных сетей.

Процессор NVIDIA Vera™ — один из ключевых

Статья получена 20 марта 2026.
Д.Е. Намиот – МГУ имени М.В. Ломоносова (email: dnamiot@gmail.com)
В.А. Сухомлин – МГУ имени М.В. Ломоносова (email: sukhomlin@mail.ru)

[1] Платформа названа в честь американского астронома Веры Флоренс Купер Рубин, известной своими работами по исследованию темной материи

[2] https://www.techpowerup.com/gpu-specs/nvidia-rubin.g1106

[3] https://www.buildmvpfast.com/blog/nvidia-vera-rubin-h300-agentic-ai-inference-cost-2026

элементов платформы. Это специализированный Arm-чип для агентных ИИ-нагрузок и крупных ИИ-фабрик. Чип содержит 88 ядер Armv9.2 Olympus с поддержкой пространственной многопоточности.

Пространственная многопоточность (она же одновременная многопоточность или SMT - Simultaneous Multithreading/Spacial Multithreading) есть один из основных подходов к многопоточности, применяемый в современных высокопроизводительных процессорах для вывода ИИ [1].

В отличие от традиционной временной многопоточности (разделение по времени), пространственная многопоточность физически разделяет аппаратные ресурсы одного ядра для одновременного выполнения нескольких задач.

Ключевые особенности поддержки пространственной многопоточности:

- Физическое разделение, когда вместо переключения между потоками на основе времени (временная многопоточность), пространственная многопоточность назначает разные аппаратные подсистемы процессора разным потокам.
- Повышенная производительность и эффективность, которая обеспечивает стабильную и предсказуемую производительность для рабочих нагрузок, требующих высокой пропускной способности, таких как перемещение кэша ключ-значение при инференсе.
- Снижение конкуренции за ресурсы в силу того, что потоки распределяются пространственно, а не конкурируют за одни и те же ресурсы с течением времени, то это снижает конфликт из-за кэш-линий.

Процессор Vera (ядра Olympus) поддерживает 88 ядер и 176 потоков, что позволяет использовать два потока на ядро за счет пространственного распределения.

Система работает с объемом памяти до 1,5 ТБ LPDDR5x, обеспечивая пропускную способность до 1,2 ТБ/с. При этом связка с GPU реализована через интерфейс NVLink, который обеспечивает 1,8 ТБ/с в дуплексном режиме.

Для обеспечения требований к задержке в реальном времени при обслуживании современных LLM и для максимально возможного количества пользователей необходимы многопроцессорные вычисления. Низкая задержка улучшает пользовательский опыт. Высокая пропускная способность снижает стоимость обслуживания. Оба фактора важны одновременно.

Даже если большая модель может поместиться в памяти одной современной графической карты, скорость генерации токенов этой картой зависит от общего объёма вычислительных ресурсов, доступных для обработки запросов. Соответственно, нужно объединять вычислительные возможности нескольких передовых графических карт.

Используя суммарную вычислительную производительность нескольких графических процессоров с такими методами, как тензорный параллелизм (TP) [2], для запуска больших моделей, запросы на вывод могут обрабатываться достаточно быстро, чтобы обеспечить ответы в реальном времени. При этом вывод с использованием нескольких графических процессоров требует больших объемов данных.

Вывод с использованием нескольких графических процессоров с тензорным параллелизмом работает путем разделения вычислений каждого слоя модели между двумя, четырьмя или даже восемью графическими процессорами на сервере. Теоретически, две видеокарты могут обрабатывать модель в 2 раза быстрее, четыре - в 4 раза быстрее, а восемь - в 8 раз быстрее[4].

Однако каждая видеокарта не может выполнять свою работу независимо. После завершения выполнения своей части слоя модели каждая видеокарта должна отправить результаты вычислений каждой другой видеокарте (связь каждый с каждым). Только после этого выполнение вывода может перейти к следующему слою модели. Соответственно, минимизация времени, затрачиваемого на передачу результатов между видеокартами, имеет решающее значение, поскольку во время этой передачи ядра тензоров часто простаивают, ожидая продолжения обработки данных.

По данным NVIDIA, один запрос к Llama 3.1 70B (8K входных токенов и 256 выходных токенов) требует передачи до 20 ГБ данных синхронизации TP с каждой видеокарты. Поскольку несколько запросов обрабатываются параллельно с помощью пакетной обработки для повышения пропускной способности вывода, объем передаваемых данных увеличивается в несколько раз.

Для хорошего масштабирования на нескольких графических процессорах сервер ИИ в первую очередь должен иметь графические процессоры с отличной пропускной способностью межпроцессорных соединений на каждый графический процессор. Он также должен обеспечивать быстрое соединение, позволяющее всем графическим процессорам как можно быстрее обмениваться данными со всеми другими графическими процессорами.

Вот почему высокоскоростное соединение между графическими процессорами имеет важное значение для многопроцессорного инференса. На рисунке 1 представлено использование NVLink (источник: NVIDIA). NVSwitch критически важен для быстрого многопроцессорного инференса LLM.

При этом пиковая скорость не зависит от количества обменивающихся данными графических процессоров. То есть, NVSwitch является неблокирующим. Без

выделенных коммутаторов, в конфигурации «точка-точка», несмотря на более низкую стоимость системы, каждый графический процессор должен разделить ту же скорость соединения все выделенные соединения. То есть, фактическая скорость соединения (обмена данными) зависела бы от количества взаимодействующих графических процессоров.

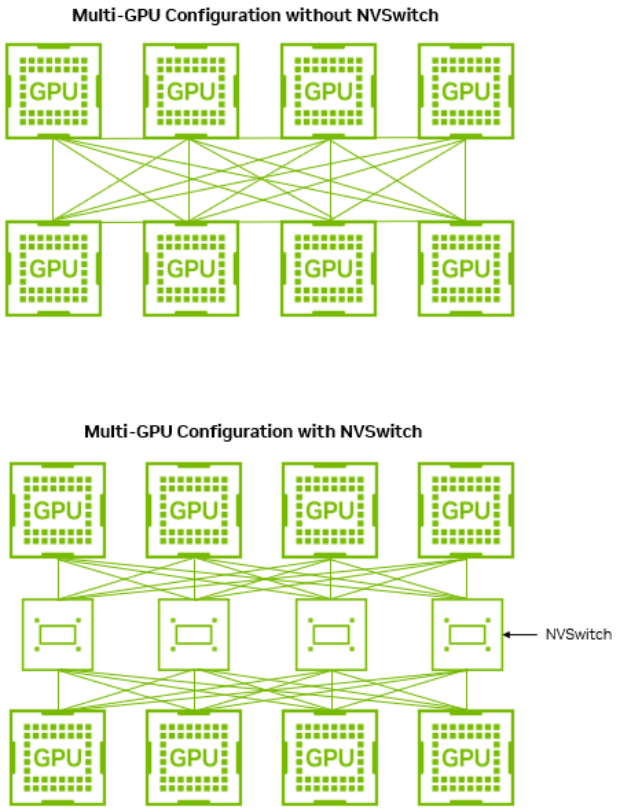

Рис. 1. NVSwitch

В отличие от простых сетевых карт, SmartNIC (NIC – Network Interface Card) допускает загрузку в контроллер дополнительного программного обеспечения уже самим пользователем. Сетевые адаптеры NVIDIA SuperNIC предлагают расширенные возможности программируемого ввода-вывода, которые имеют решающее значение для современных центров обработки данных с искусственным интеллектом. Эти карты оснащены ускоренным конвейером обработки пакетов, способным работать на скорости линии связи с пропускной способностью до 800 Гбит/с.

Перенося задачи обработки пакетов с CPU на SuperNIC, этот конвейер значительно снижает задержку в сети и повышает общую эффективность системы. Программируемый характер конвейера, работающий на основе программной платформы NVIDIA DOCA, предоставляет сетевым специалистам гибкость в построении и оптимизации сетей в масштабах крупных предприятий.

Сопроцессоры для обработки данных (DPU) — типичное расширение сетевых плат SmartNIC, к которым добавляют функциональность NVMe. NVMe (Non-Volatile Memory Express) - протокол, разработанный для использования шины PCI Express для подключения твердотельных накопителей (SSD) к серверам или процессорам. Такая плата позволяет разгрузить центральный процессор, забрав себе все задачи ввода-вывода. Сопроцессор для обработки данных освободит серверные процессоры от инфраструктурных задач. Исследования показывают, что в сильно виртуализированных средах сетевые процессы могут занимать более 30% процессорного времени на хосте. Представьте, что дисковые операции, шифрование, сканирование трафика (DPI) и сложная маршрутизация выполняются отдельным модулем. Это потенциально снимет значительную часть нагрузки с CPU.

На рисунке 2 представлена типовая архитектура DPU

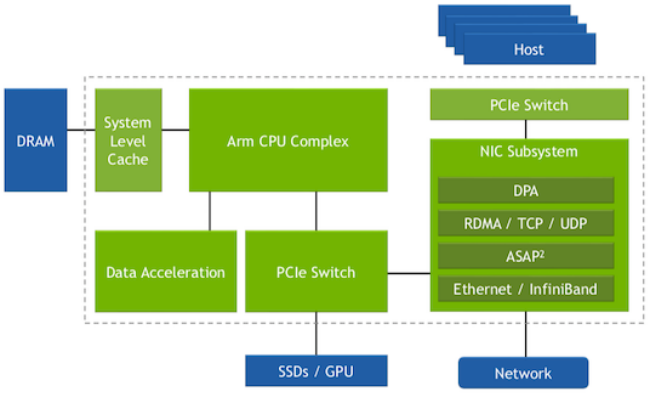

Рис.2 BlueField DPU[5]

Используя программное обеспечение NVIDIA DOCA и SPDK (Storage Performance Development Kit), DPU BlueField могут эмулировать устройства NVMe, позволяя им отображаться для хоста как локальные физические NVMe-накопители, одновременно управляя данными, шифрованием и подключением в фоновом режиме.

Ключевые приложения DPU:

- Виртуальные и аппаратные облачные среды.
- Хранилища NVMe в виртуальных машинах.
- Приложения Network Function Virtualization (NFV).
- Приложения ИБ, такие как Deep Packet Inspection (DPI).
- Микросерверы для граничных вычислений

III. Уникальные особенности

Выход платформы NVIDIA Rubin доказывает, что вывод (inference) теперь является системной проблемой, а не проблемой конкретного чипа. Отдельный «чип» больше

[4] https://developer.nvidia.com/blog/nvidia-nvlink-and-nvidia-nvswitch-supercharge-large-language-model-inference/

[5] https://chromewebstore.google.com/detail/hoxx-vpn-proxy/nbcojefnccbanplpoffopkoepjmhgdgh

не является ограничивающим фактором. Ограничивающим фактором является обеспечение чипа необходимыми ресурсами.

Из выступления CEO NVIDIA: «Будущее — это координация множества отличных моделей на каждом этапе цепочки рассуждений».

Тогда как традиционно оценивалась производительность в FLOPs, платформа Rubin атакует совсем другое узкое место. Производительность вывода больших моделей сейчас ограничена пропускной способностью памяти и пропускной способностью сети.

Современные задачи искусственного интеллекта, включая рассуждения, смешанные группы экспертов (MoE - архитектура нейронной сети, повышающая эффективность и масштабируемость ИИ за счет использования множества специализированных подсетей или «экспертов», вместо одной большой, плотной модели [3]), вывод в длинном контексте и обучение с подкреплением, не ограничиваются только пиковой производительностью операций с плавающей запятой (FLOPS). Задачи ИИ ограничивает возможность поддержания эффективности выполнения на уровне вычислительных ресурсов, памяти и связи. Графический процессор Rubin специально разработан для решения этой задачи, оптимизируя весь путь выполнения, преобразуя энергопотребление, пропускную способность и память.

Для качественного инференса, ценность системы измеряется ее «циклами рассуждений», ее способностью автономно вызывать инструменты и ее способностью управлять огромными, постоянно доступными контекстными данными.

Платформа NVIDIA Vera Rubin NVL72 объединяет перечисленные выше компоненты - 72 графических процессора Rubin, 36 центральных процессоров Vera, сетевые карты ConnectX-9 SuperNIC и DPU BlueField-4. Она масштабирует интеллектуальные возможности на стоечной платформе с помощью коммутатора NVIDIA NVLink 6 и расширяет их с помощью NVIDIA Quantum-X800 InfiniBand и Spectrum-X Ethernet, обеспечивая масштабируемую промышленную революцию в области ИИ. Платформа обеспечивает новый класс производительности при выполнении инференции для моделей с триллионами параметров и контекстом в миллионы токенов.

Типичные агентные рабочие нагрузки сегодня используют идентичные большие контекстные окна для множества запросов. Для достижения такой производительности необходимо использовать кэширование ключ-значение (KV-кэш). Выполнение этой операции на одном узле неэффективно, поэтому используется распределенное кэширование ключ-значение. В маркетинге NVIDIA это называется NVIDIA Inference Context Memory Storage Platform (ICMSP - платформа хранения контекста и памяти для вывода). Платформа ICMSP, работающая на базе BlueField-4, представляет собой выделенный уровень хранения, предназначенный для управления и совместного использования кэша «ключ-значение» (KV) для крупномасштабного вывода ИИ с длительным контекстом. Она использует CMX (Context Memory Storage) для перемещения данных KV между высокоскоростным хранилищем и памятью графического процессора, обеспечивая до 5 раз более высокую скорость обработки токенов в секунду (TPS) и большую эффективность по сравнению с традиционными хранилищами.

Для максимальной пропускной способности необходимо использовать пакеты данных очень большого размера. Это требует разделения модели на несколько узлов и быстрой сети между ними.

Для обучения необходимо согласовывать веса после нескольких шагов. Это нетривиальная проблема, поскольку трафик очень импульсный, а задержка в хвосте распределения влияет на производительность.

Чтобы понять, почему платформа Рубина принципиально отличается от предшествующих ей, необходимо понять уникальные аппаратные требования к агенту ИИ. Агент планирует, действует и проверяет [4]. Например, агент должен:

Планировать: разбить задачу на подзадачи.

Действовать: открыть браузер, выполнить поиск, перейти к базе данных и запустить скрипт для анализа данных.

Проверять: проанализировать результаты в сравнении с первоначальной целью.

Исправить: повторить поиск, если исходных данных недостаточно.

Этот «цикл» является ядром агентного ИИ [5]. Для аппаратного обеспечения это создает проблему. Стандартные графические процессоры (GPU) разработаны для параллельной обработки больших объемов данных, что идеально подходят для обучения модели. Но агенту, постоянно находящемуся в «цикле мышления», требуется крайне низкая задержка при выводе результатов, быстрый доступ к памяти для вызова инструментов и высокая производительность однопоточного ЦП для управления программной средой («песочницей»), в которой работает агент [6].

Платформа Rubin от NVIDIA — это первый «полноценный» ответ на эти требования к агентам. В то время как предыдущие поколения были, в первую очередь, графическими процессорами, Rubin – это, в первую очередь, система, состоящая из взаимосвязанных компонент, разработанных для функционирования в качестве единого суперкомпьютера для ИИ. Классическое решение проблемы силоса данных [7].

Для выполнения сложных рассуждений агенту ИИ необходимо поддерживать огромный контекстный

диапазон. Если агент «забывает» начало документа при составлении его заключительной части, цикл рассуждений прерывается. Rubin GPU решает эту проблему добавлением 288 Гб памяти HBM4 (High Bandwidth Memory 4) на один процессор. Модели с большим количеством параметров могут работать полностью в одной стойке NVL72 без задержек, характерных для многоузловых распределенных систем. При этом обеспечивается пропускная способность памяти 22 ТБ/с.

Сетевые карты ConnectX-9 обеспечивают пропускную способность 1,6 ТБ/с на каждый GPU. Эту огромную пропускную способность в 1,6 ТБ/с можно использовать для динамической потоковой передачи и замены экспертов (моделей). Соответственно, происходит переход от «статического вывода» (загрузка весов и ожидание) к «системной оркестровке» (управление состоянием на 72 графических процессорах в реальном времени).

Также меняется роль центрального процессора (ЦП). Это теперь не просто «хост», передающий данные на графический процессор. В Рубин CPU является оркестратором.

Агенты выполняют код. Они должны использовать «песочницу», чтобы гарантировать безопасность своих автономных действий [8,9]. Это последовательные, логически сложные задачи, с которыми не могут справиться графические процессоры. Процессор NVIDIA Vera имеет 88 пользовательских ядер «Olympus», разработанных для решения этой проблемы.

Традиционные ЦП используют многопоточность с «разделением по времени», где потоки поочередно используют ресурсы ядра. При высокой нагрузке это создает «дрожание» (скачки задержки), которое может нарушить цикл рассуждений агента. NVIDIA Vera представляет пространственную многопоточность, которая физически разделяет ресурсы ядра. Это обеспечивает детерминированную производительность: каждая среда агента (песочница) получает выделенный, изолированный блок ресурсов [10].

Пропускная способность памяти NVIDIA Vera - 1,2 ТБ/с. Vera обеспечивает в 3 раза большую пропускную способность на ядро по сравнению с традиционными процессорами для центров обработки данных, гарантируя, что ресурсоемкие задачи, такие как ETL и аналитика в реальном времени, не будут тормозить.

IV. Эмуляторы и цифровые двойники

NVIDIA также выпустила Omniverse DSX Blueprint - инструмент для цифровых двойников фабрик ИИ. Этот инструмент [6] создан для разработчиков, чтобы ускорить проектирование и эксплуатацию фабрик ИИ гигаваттного масштаба за счет интеграции физических и цифровых данных в интерактивные цифровые двойники, построенные на платформе OpenUSD [11]. OpenUSD (Universal Scene Description) - это высокопроизводительная платформа с открытым исходным кодом для описания и обмена данными в 3D-сценах, первоначально разработанная компанией Pixar. Она выступает в качестве стандартизированного языка для неразрушающего редактирования, совместной работы и беспрепятственного обмена данными между 3D-приложениями, что делает её де-факто стандартом для 3D-графики, промышленной цифровизации и обучения ИИ.

Используя ресурсы SimReady (Simulation-Ready - это стандарт NVIDIA для 3D-моделей, разработанных для реалистичного поведения в виртуальных симуляциях, выходящего за рамки простого визуального отображения. В отличие от стандартных 3D-моделей, модели SimReady создаются на основе OpenUSD и включают в себя встроенные физические свойства, поведение и метаданные) и библиотеки Omniverse (коллекция библиотек для проектирования физического ИИ), разработчики могут интегрировать моделирование энергопотребления, тепловых характеристик и эксплуатационных параметров в реальном времени непосредственно в свои рабочие процессы для повышения эффективности, экологичности и надежности.

Проект предоставляет комплексную основу для создания, моделирования и оптимизации центров обработки данных ИИ гигаваттного масштаба — плавно переходя от разработки пользовательских приложений и физически точного проектирования к высокоэффективным, устойчивым к энергоснабжению операциям.

Репозиторий проекта содержит:

1) Геометрию цифрового двойника, основанную на всей эталонной конструкции DSX для площадки площадью 0.2 км^2 (50 акров), включая вычислительное здание и вспомогательную инфраструктуру.
2) Веб-приложение с пользовательским интерфейсом, разработанным с использованием библиотек Omniverse, для взаимодействия с цифровыми двойниками, просмотра результатов моделирования, а также создания и сохранения конфигураций сборки.
3) Готовые к моделированию ресурсы для ускорения создания цифровых двойников:
4) Моделирование тепловых потоков в коридорах с использованием вычислительной гидродинамики (CFD)
5) Примеры вычислительных конфигураций
6) Моделирование электрической нагрузки для тестирования ее различных конфигураций

[6] https://docs.omniverse.nvidia.com/dsx/latest/index.html

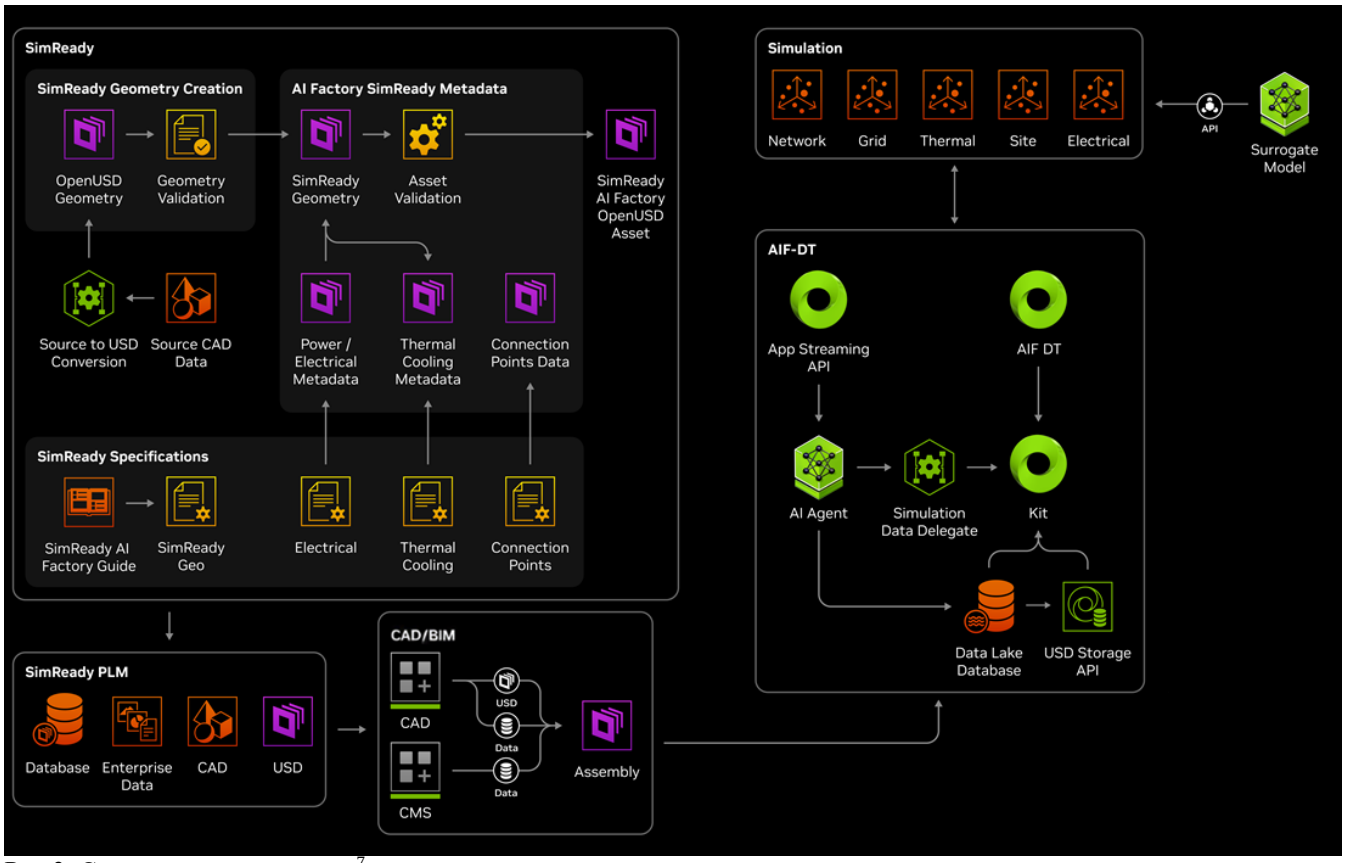

Рис.3. Система моделирования[7]

Общая архитектура показана на рис. 3.

V Заключение

Очевидно, что с выпуском платформы Vera Rubin компания NVIDA открыла новую эру в проектировании аппаратного обеспечения. В принципе, история разработки, например, процессоров, ориентированных на поддержку специализированной программной архитектуры совсем не нова. Можно вспомнить, например, Intel iAPX 432 (1981 год), система команд которого поддерживала работу со сложными структурами данных, что давало возможность сократить объём программного кода операционной системы (по сравнению с объёмом кода для процессоров с другой системой команд). Сопроцессоры, как специализированные вычислители, также используются уже достаточно долго. Платформа Vear Rubin являет же собой пример согласованной архитектуры для решения целого класса задач, а не просто набор специализированных компонент. Можно предположить, что мы увидим развитие такого подхода для других классов задач, например, для конкретных типов ИИ-агентов.

Выпуск платформы, очевидно, подстегнет интерес к средствам моделирования компьютерной архитектуры. При этом интересно будет именно моделирование и настройка программного обеспечения на новых проектируемых системах.

Библиография

[1] Hsu, Kuan-Chieh, and Hung-Wei Tseng. "Simultaneous and heterogenous multithreading: Exploiting simultaneous and heterogeneous parallelism in accelerator-rich architectures." IEEE Micro 44.4 (2024): 11-19.

[2] Li, Qingyuan, et al. "Flash communication: Reducing tensor parallelization bottleneck for fast large language model inference." arXiv preprint arXiv:2412.04964 (2024).

[3] Mu, Siyuan, and Sen Lin. "A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications." arXiv preprint arXiv:2503.07137 (2025).

[4] Namiot, Dmitry, and Eugene Ilyushin. "On Architecture of LLM agents." International Journal of Open Information Technologies 13.1 (2025): 67-74.

[5] Durante, Zane, et al. "Agent ai: Surveying the horizons of multimodal interaction." arXiv preprint arXiv:2401.03568 (2024).

[6] Krishnan, Naveen. "Ai agents: Evolution, architecture, and real-world applications." arXiv preprint arXiv:2503.12687 (2025).

[7] Цифровая экономика и Интернет Вещей - преодоление силоса данных / В. П. Куприяновский, А. Р. Ишмуратов, [и др.] // International Journal of Open Information Technologies. – 2016. – Т. 4, № 8. – С. 36-42. – EDN WFVAPB.

[8] Namiot, Dmitry, and Eugene Ilyushin. "On the Cybersecurity of AI Agents." International Journal of Open Information Technologies 13.9 (2025): 13-24.

[9] Buscemi, Alessio, et al. "Towards Sandboxes for the Internet of Agents." Available at SSRN 5801322 (2025).

[10] Zheng, Yusheng, et al. "AgentCgroup: Understanding and Controlling OS Resources of AI Agents." arXiv preprint arXiv:2602.09345 (2026).

[11] Ahn, Seok-hyun, et al. "Studying the Universal Scene Description (USD) file format from a Digital Twin Convergence Perspective." International journal of advanced smart convergence (2025): 224-241.

[7] https://docs.omniverse.nvidia.com/dsx/latest/system-architecture.html#system-architecture

NVIDIA Vera Rubin as a new standard for artificial intelligence infrastructure

Dmitry Namiot, Vladimir Sukhomlin

***Abstract* - This article analyzes the new NVIDIA Vera Rubin platform, introduced in 2026 and positioned by the company as a qualitative leap in building computing infrastructure for artificial intelligence. Unlike traditional approaches focused on individual chips, the platform is considered as a holistic system, combining Vera (Arm) processors, Rubin graphics accelerators, high-speed NVLink interfaces, ConnectX-9 network adapters, and programmable BlueField-4 DPUs. Particular attention is paid to hardware support for agent-based AI, including spatial multithreading, distributed key-value caching, and the scalable NVL72 rack-mount architecture. A separate section is devoted to the use of digital twins based on Omniverse DSX for the design and operation of large-scale AI factories. The authors conclude that the Vera Rubin platform marks a shift from performance evaluation by peak FLOPS to system-wide optimization of memory and network bandwidth, setting new standards for infrastructure solutions in the field of artificial intelligence.**

***Keywords*- NVIDIA, Artificial Intelligence, Vera Rubin.**

REFERENCES

[1]Hsu, Kuan-Chieh, and Hung-Wei Tseng. "Simultaneous and heterogenous multithreading: Exploiting simultaneous and heterogeneous parallelism in accelerator-rich architectures." IEEE Micro 44.4 (2024): 11-19.

[2]Li, Qingyuan, et al. "Flash communication: Reducing tensor parallelization bottleneck for fast large language model inference." arXiv preprint arXiv:2412.04964 (2024).

[3]Mu, Siyuan, and Sen Lin. "A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications." arXiv preprint arXiv:2503.07137 (2025).

[4]Namiot, Dmitry, and Eugene Ilyushin. "On Architecture of LLM agents." International Journal of Open Information Technologies 13.1 (2025): 67-74.

[5]Durante, Zane, et al. "Agent ai: Surveying the horizons of multimodal interaction." arXiv preprint arXiv:2401.03568 (2024).

[6]Krishnan, Naveen. "Ai agents: Evolution, architecture, and real-world applications." arXiv preprint arXiv:2503.12687 (2025).

[7]Kupriyanovsky, Vasily, et al. "Digital Economy and the Internet of Things-negotiating data silo." International Journal of Open Information Technologies 4.8 (2016): 36-42.

[8]Namiot, Dmitry, and Eugene Ilyushin. "On the Cybersecurity of AI Agents." International Journal of Open Information Technologies 13.9 (2025): 13-24.

[9]Buscemi, Alessio, et al. "Towards Sandboxes for the Internet of Agents." Available at SSRN 5801322 (2025).

[10] Zheng, Yusheng, et al. "AgentCgroup: Understanding and Controlling OS Resources of AI Agents." arXiv preprint arXiv:2602.09345 (2026).

[11] Ahn, Seok-hyun, et al. "Studying the Universal Scene Description (USD) file format from a Digital Twin Convergence Perspective." International journal of advanced smart convergence (2025): 224-241.