

Keysight AI Inference Builder

Валидация, бенчмаркинг и оптимизация инфраструктуры инференса ИИ с помощью эмуляции рабочих нагрузок ИИ

Введение

Keysight AI Inference Builder (KAI IB) — решение для эмуляции и аналитики, предназначенное для валидации, бенчмаркинга и оптимизации инфраструктур инференса ИИ и программных стеков. Решение эмулирует реалистичные рабочие нагрузки ИИ с высокой точностью и в большом масштабе, предоставляя детальную информацию о производительности, возможностях и эффективности механизмов безопасности инференс-систем.

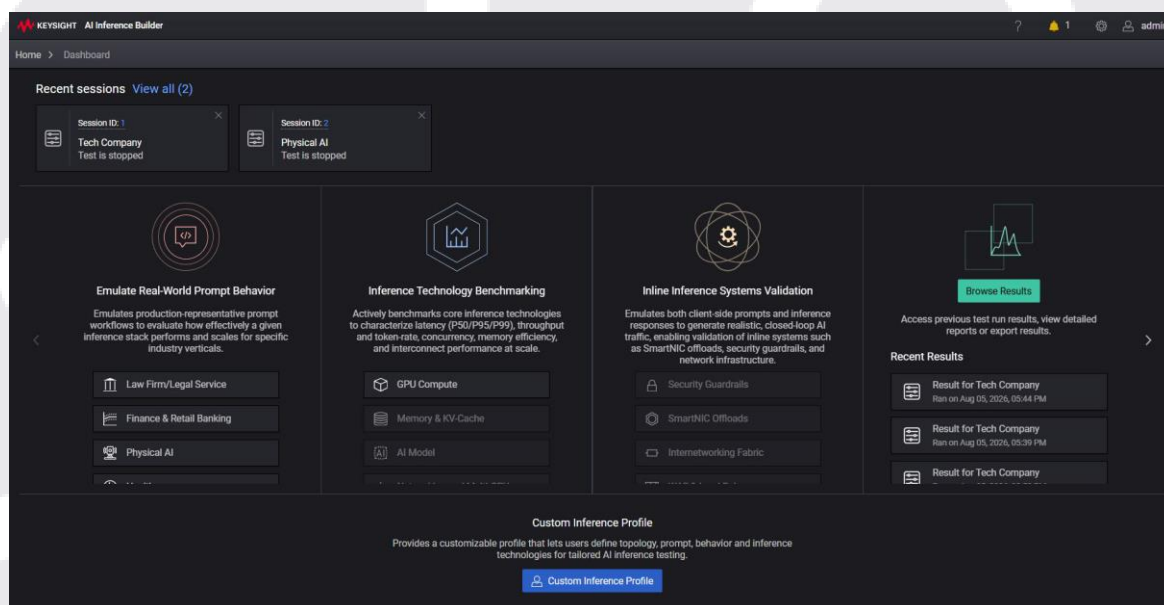


Рисунок 1. Главная страница пользовательского интерфейса KAI Inference Builder.

Основные возможности

- Эмуляция реалистичного поведения клиентов ИИ в большом масштабе для проверки всей инфраструктуры и программного стека инференса.
- Выбор различных профилей пользовательских запросов ИИ, создающих нагрузку на разные этапы цепочки инференса.
- Проверка инфраструктуры инференса ИИ, развёрнутой в публичном или частном облаке, с использованием полностью виртуальной либо аппаратной эмуляции клиентов инференса.
- Масштабирование до миллионов пользователей с детальным управлением интенсивностью запросов в секунду.
- Доступ к подробной статистике инференса для выявления узких мест, ограничений и неэффективности различных компонентов цепочки инференса, включая:
 - GPU-вычисления
 - память HBM / VRAM

Тел: +7 (495) 252-00-96

Email: info-site@akmetron.ru

Адрес: Москва, ул. Рабочая, д.93 стр.2, под. 2.

- KV-cache
- движки моделей и оркестраторы
- Корреляция клиентских метрик с телеметрией движка инференса (например, статистикой vLLM) и системной GPU-телеметрией (например, данными DCGM) в едином синхронизированном по времени представлении:
 - запросы в секунду
 - одновременные запросы
 - Time to First Token (TTFT) — максимум и процентиля, например P50, P90, P99
 - Time to Last Token (TTLT) — максимум и процентиля, например P50, P90, P99
 - токены в секунду — входные / выходные
 - использование кэша
 - время prefill и decode
 - использование Tensor Core
 - состояние планировщика
 - энергопотребление GPU
- Повторяемые нагрузочные, стрессовые, устойчивостные и robustness-тесты инфраструктур и стеков инференса с автоматизированными сценариями как короткой продолжительности, так и длительными soak-тестами.

Эмуляция рабочих нагрузок KAI Inference Builder

KAI Inference Builder формирует реалистичные пользовательские запросы из разных отраслей — финансовых услуг, здравоохранения, разработки ПО и программирования, юридических услуг — для точной оценки производительности инфраструктуры инференса ИИ.

Подходы, основанные только на синтетическом HTTP-трафике, не воспроизводят реальное поведение инференса и могут скрывать узкие места. KAI IB использует глубоко проработанную эмуляцию клиентских запросов, детальное управление частотой запросов и масштабом пользователей, а также предоставляет KPI, специфичные для инференса.

Реалистичная генерация запросов позволяет создавать full-loop трафик инференса и нагружать всю инфраструктурную цепочку:

- Front-end сетевые решения, например балансировщики нагрузки, прокси и Web Application Firewall (WAF)
- шлюзы безопасности ИИ и сетевые механизмы guardrails
- SmartNIC / DPU
- серверы инференса ИИ и программные стеки: GPU, HBM / VRAM, KV-cache и движки инференса

Тел: +7 (495) 252-00-96

Email: info-site@akmetron.ru

Адрес: Москва, ул. Рабочая, д.93 стр.2, под. 2.



Рисунок 2. Эмуляция KAI IB проверяет всю цепочку инфраструктуры инференса ИИ.

Функциональные возможности KAI Inference Builder

Решение KAI IB состоит из одного или нескольких агентов трафика, генерирующих реальные запросы инференса, и одного контроллера с веб-интерфейсом для управления сессиями, настройки и запуска тестов, а также сбора актуальной статистики.

Агенты трафика KAI IB

- Использование специализированных аппаратных конечных точек с высокой производительностью либо лёгких программных тестовых агентов, независимых от инфраструктуры и пригодных для публичных облаков и локальных виртуальных машин.
- Генерация реалистичных запросов ИИ, направленных в реальный стек инференса и проходящих через всю инфраструктуру публичного или частного облака.
- Выбор разных пользовательских персон ИИ с профилями запросов для финансовой отрасли, здравоохранения, разработки / программирования и юридических услуг, а также технологических бенчмарков, например GPU compute.
- Масштабирование до миллионов моделируемых пользователей для нагрузочного тестирования в производственном масштабе.
- Точное управление числом запросов в секунду для поиска узких мест инфраструктуры, ограничений GPU-вычислений, пропускной способности памяти и неоптимальной работы движков инференса.

Контроллер KAI IB

- Полностью cloud-native приложение на основе микросервисов с эластичным масштабированием; разворачивается как виртуальная машина локально или в публичном облаке. Архитектура на базе Kubernetes обеспечивает масштабируемость, отказоустойчивость и самовосстановление.
- Современный веб-интерфейс позволяет настраивать и запускать тесты без отдельного тяжёлого клиентского приложения.

Тел: +7 (495) 252-00-96

Email: info-site@akmetron.ru

Адрес: Москва, ул. Рабочая, д.93 стр.2, под. 2.

- Интерфейс с поддержкой сессий и многопользовательской аутентификации рассчитан на индивидуальную и совместную работу команд: управление сессиями, загрузка конфигураций, запуск тестов и мониторинг результатов.
- Представление KPI со стороны клиента: запросы в секунду, Time to First Token, Time to Last Token и число одновременных запросов.
- Единое представление статистики объединяет данные клиента с серверными метриками инференса и GPU: скорость токенов, prefill/decode, использование кэша, энергопотребление GPU и использование Tensor Core.
- Расширенная диагностика с детальной CSV-статистикой уровней L2-L7 и захватом пакетов на портах.
- REST API позволяет интегрировать KAI IB в современные системы автоматизации: настраивать тесты, эмулировать приложения и получать результаты через API.

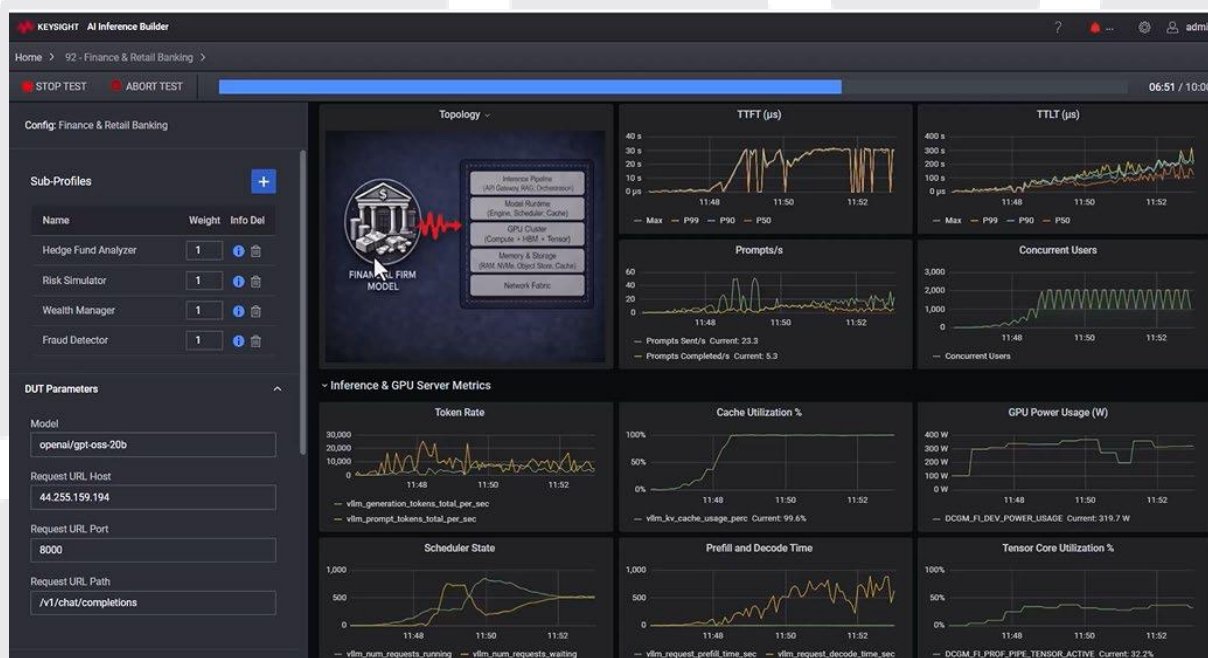


Рисунок 3. Единое представление статистики KAI IB.

Аппаратные платформы KAI Inference Builder

Семейство платформ Keysight APS-100 / 400GE следующего поколения для тестирования приложений и безопасности включает контроллер APS-M1010 и устройства APS-ONE-100 / APS-ONE-400. Архитектура позволяет поэтапно наращивать систему. В режиме stack устройства APS-ONE-100 или APS-ONE-400 управляются контроллером APS-M1010, который поддерживает до десяти устройств в одной системе.

Вычислительный узел Keysight AI Inference Builder, например 947-4090, основан на APS-ONE-400 и предназначен исключительно для работы ПО KAI IB. Управление осуществляется вычислительным узлом APS-M1010. Такая конфигурация предназначена для высокопроизводительных тестовых сценариев и сложных вариантов развёртывания.

Тел: +7 (495) 252-00-96

Email: info-site@akmetron.ru

Адрес: Москва, ул. Рабочая, д.93 стр.2, под. 2.

Технические характеристики

Ключевая характеристика	Параметры
Варианты развёртывания агентов	Публичные облака: Amazon Web Services (AWS). Частные облака / локальное развёртывание: VMware ESXi 8.0, 7.0 и ESXi 6.X.
Варианты развёртывания контроллера	Публичные облака: Amazon Web Services (AWS). Частные облака / локальное развёртывание: VMware ESXi 8.0, 7.0 и ESXi 6.X.
Поддерживаемые конфигурации SUT	Квалифицированный движок инференса: vLLM. Формат API запросов: OpenAI API — chat completions. Квалифицированные LLM-модели: GPT-OSS, Llama, Qwen, Mistral, Gemma.
Поддерживаемые типы целевой нагрузки	Запросы в секунду. Моделируемые пользователи — вторичный тип.
Поддерживаемые типы запросов	Реалистичное поведение запросов: финансовые услуги, юридические услуги, здравоохранение, tech / coding, Physical AI. Бенчмаркинг технологии инференса: GPU compute.
Ключевые показатели производительности	Одновременные запросы. Time to First / Last Token — максимум и процентиля, например P50, P90, P99. Длина входной / выходной последовательности в токенах — процентиля P50, P90, P99. Запросы в секунду.
Статистика инференса и GPU-сервера	Скорость токенов. Время prefill и decode. Использование кэша. Состояние планировщика. Энергопотребление GPU. Использование Tensor Core.
Отчётность	PDF. Подробная CSV-статистика уровней L2-L7. Захват пакетов на портах.
Максимальное число IP-адресов на агент	10 000.

Тел: +7 (495) 252-00-96

Email: info-site@akmetron.ru

Адрес: Москва, ул. Рабочая, д.93 стр.2, под. 2.

Автоматизация	Полное покрытие REST API. Документация REST API.
---------------	---

Информация для заказа — программные комплекты

Номер	Описание
952-1010	KAI Inference Builder Bundle с 10 агентами и до 1000 запросов в секунду — подписка на 1 год, плавающая мировая лицензия, TAA Compliant. Комплект AI Inference Performance and Security Testing включает: • доступ к cloud-native приложению KAI Inference Builder; • до 10 тестовых агентов KAI IB, разворачиваемых в среде заказчика; • один performance unit до 1000 запросов в секунду и 10 000 одновременных пользователей, который может быть использован в одном параллельном тесте; • доступ к API, обновлениям ПО и поддержке на оплаченный срок подписки. Требуется указать срок лицензии; приобретается кратно целому числу лет. Цена указывается за единицу в год.
952-1100	KAI Inference Builder Bundle с 10 агентами и до 10 000 запросов в секунду — подписка на 1 год, плавающая мировая лицензия, TAA Compliant. Комплект включает: • доступ к cloud-native приложению KAI Inference Builder; • до 10 тестовых агентов KAI IB в среде заказчика; • один performance unit до 10 000 запросов в секунду и 100 000 моделируемых пользователей, который может быть использован в одном параллельном тесте; • доступ к API, обновлениям ПО и поддержке на оплаченный срок подписки. Требуется указать срок лицензии; приобретается кратно целому числу лет. Цена указывается за единицу в год.

Информация для заказа — аппаратная часть

Номер	Описание
950-2400	Бессрочная лицензия KAI Inference Builder для одного вычислительного узла , плавающая лицензия на контроллер. Активирует KAI Inference Builder на

Тел: +7 (495) 252-00-96

Email: info-site@akmetron.ru

Адрес: Москва, ул. Рабочая, д.93 стр.2, под. 2.

	одном вычислительном узле APS-ONE-100, APS-ONE-400 или KAI Inference Builder Compute Node. Может использоваться с узлом, управляемым контроллером APS-M1010. Соответствие ТАА определяется аппаратной платформой.
941-0116	APS-M1010 Management Controller. Поддерживает приложения BreakingPoint, CyPerf, KAI IB и IxLoad. Может управлять до 10 узлов APS-ONE-100 / APS-ONE-400 / KAI IB Compute Node (941-0110, 941-0114, 947-4080, 947-4077, 947-4090). Вычислительные узлы приобретаются отдельно.
947-4090	KAI Inference Builder Compute Node с 4 × 100GE. Совместим с контроллером APS-M1010. Требуется отдельного заказа трансиверов / кабельных сборок QSFP28: QSFP28-SR4-XCVR (100GBASE-SR4, MMF, 850 нм, до 100 м) либо 942-0088 (пассивный медный Direct Attach Cable). Также требуется бессрочная лицензия KAI Inference Builder 950-2400 для одного вычислительного узла; лицензия приобретается отдельно одновременно с оборудованием.

Тел: +7 (495) 252-00-96

Email: info-site@akmetron.ru

Адрес: Москва, ул. Рабочая, д.93 стр.2, под. 2.