Keysight AI Inference Builder

Validate, Benchmark, and Optimize AI Inference Infrastructures with AI Workload Emulation

Introduction

The Keysight AI Inference Builder (KAI IB) is an emulation and analytics solution designed to validate, benchmark and optimize AI inference infrastructures and software stacks by emulating realistic AI workloads with high fidelity and at scale providing deep insights into the performance characteristics, capabilities and security efficacy of inference systems.

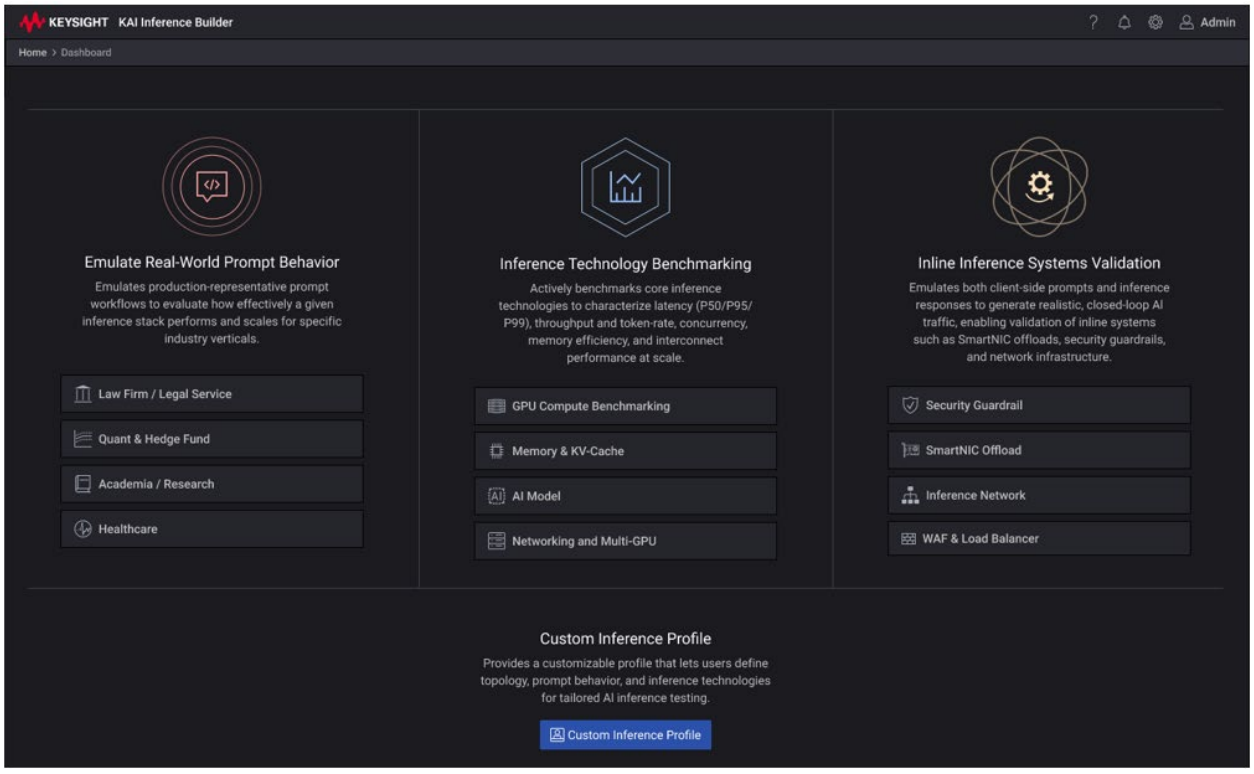

Figure 1. KAI Inference Builder UI landing page

KEYSIGHT

Highlights

- Emulate realistic AI client behavior at scale to validate entire AI inference infrastructures and stacks.
- Choose different AI persona prompts, driving pressure points at different stages of the AI inference layers.
- Validate public or private cloud-deployed AI inference infrastructures with fully virtual or hardware base inference client emulation.
- Scale to emulate millions of users with granular control of the generated prompts-per-second load for unmatched AI inference scale testing.
- Access detailed inference statistics to gain actionable insights into potential bottlenecks, limits or inefficiencies at various components of the AI inference pipeline, such as:
  - GPU compute
  - HBM / VRAM memory systems
  - KV-cache and storage layers
  - Model engines and orchestrators
- Correlate client-side metrics with the ingestion of inference engine level telemetry (for example, VLLM statistics), and system-level GPU telemetry (for example, DCGM data) in a single time-synchronized view:
  - Prompts ser second
  - Concurrent Users
  - Time to First Token (TTFT) — Max and percentiles (for example, P50, P90, P99)
  - Time to Last Token (TTLT) — Max and percentiles (for example, P50, P90, P99)
  - Tokens per second (input / output)
  - Cache Utilization
  - Prefill and Decode Time
  - Tensor Core Utilization
  - Scheduler State
  - GPU Power Usage
- Perform a resilience and robustness testing of AI inference infrastructures and stacks with automated test scenarios for repetitive short duration or long duration soak tests.

KAI Inference Builder Workload Emulation

KAI Inference Builder (KAI IB) delivers new heights in realism by generating real-world user prompts across verticals such as financial services and law / legal services to accurately characterize the performance of AI inference infrastructures.

Other inference test approaches using synthetic HTTP traffic fail to reproduce real inference behavior and can hide bottlenecks, creating costly surprises in production.

KAI IB aims to close this gap with deeply researched AI inference client prompt emulation, granular control of the prompt rate and user scale as well as providing actionable insights with inference-native KPIs.

Highly realistic prompt generation allows KAI IB to generate inference traffic that flows in a full-loop approach, stressing and validating the entire AI inference infrastructure chain, such as:

- Front-end network solutions (for example, load balancers, proxies, web application firewalls)
- AI security gateways and guardrails network controls
- SmartNICs / DPUs
- AI inference servers and software stacks (GPU, HBM / VRAM, KV-cache and inference engines)

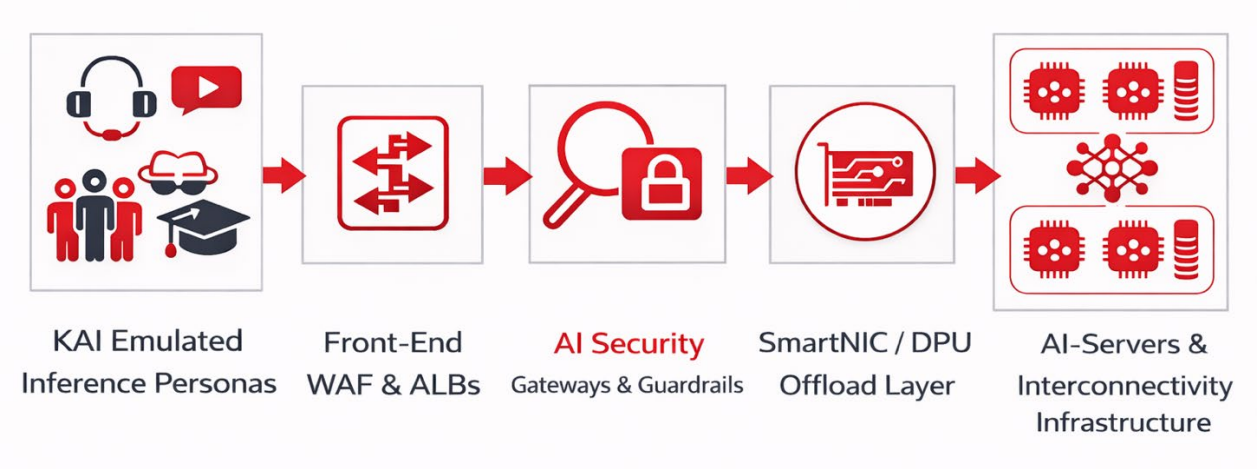

Figure 2. KAI IB emulation validates the entire AI inference infrastructure chain

KAI Inference Builder Features

The KAI IB solution is comprised of one or multiple traffic agents (the entities generating the actual inference prompt traffic) and one controller (web-based UI for session management, configuring and running tests, as well as capturing relevant real-time statistics).

KAI IB Traffic Agents

- Use hardware dedicated endpoints with unparalleled performance or deploy lightweight software-based test agents that are infrastructure-agnostic, allowing operations on public cloud instances or on-prem Virtual Machines (VMs)
- Generate realistic AI prompts targeted at a real AI inference stack that can traverse an entire inference infrastructure deployed over public or private clouds
- Select different AI client user personas with different prompt profiles relevant for various verticals (for example, financial, legal) or technology benchmarks (for example, GPU compute)
- Scale up to millions of simulated users to assess the AI inference infrastructure and software stack under production-scale load testing
- Granular control over the number of prompts per second to identify infrastructure bottlenecks, GPU compute or memory bandwidth limits or sub-optimal operation of inference engines

KAI IB Controller

- KAI IB controller is a completely cloud-native, micro-services-based, elastically scalable application deployed as a virtual machine (on-prem or in public clouds). Developed on top of Kubernetes-based architecture, it leverages attributes that make it scalable, resilient, and self-healing.
- Features a modern, easy-to-use web-based User Interface (UI) making it flexible to configure and run tests without introducing the friction of a thick dedicated application.
- The session-aware UI supports multi-user authentication. Session support is tailor-made for teams to manage sessions individually or collaboratively, upload configurations, run tests, and monitor results.
- Detailed inference-native KPIs are presented from a client perspective, such as prompts per second, time to first token, time to last token, concurrent simulated users.
- Single-pane-of-glass statistics view enables unparalleled collaboration across different teams with the ingestion of the server-side inference and GPU server metrics such as token rate, prefill and decode time, cache utilization, GPU power usage, and tensor core utilization.
- Developed from the ground up with a REST API approach, it enables integration of KAI IB into modern automation frameworks where users can configure tests, emulate applications, and gather results — all through REST API calls.

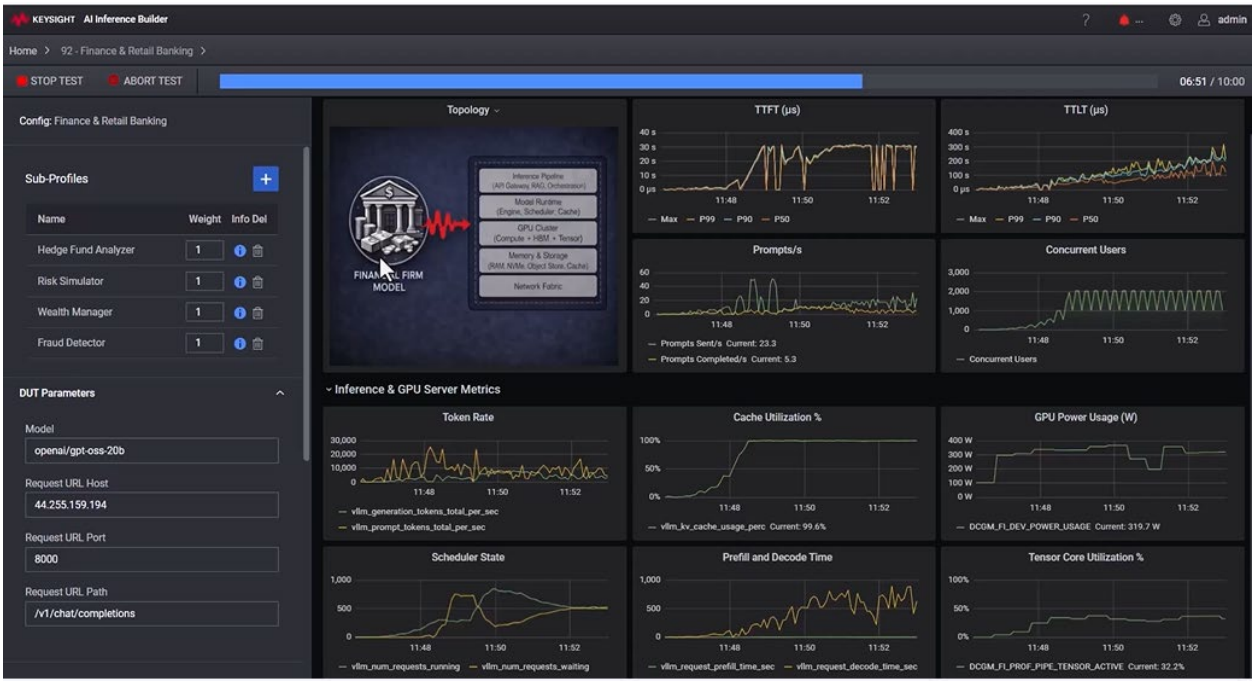

Figure 3. KAI IB single-pane-of-glass statistics view

KAI Inference Builder Hardware Platforms

The Keysight APS-100 / 400GE family next-generation application and security test platform, consisting of the APS-M1010 controller and APS-ONE-100 appliance, recreates hyperscale environments in a modular, grow-as-you-need approach. With KAI IB, the APS-ONE-100 appliance is used in a stack mode and controlled by the APS-M1010 controller, which supports up to ten appliances in a single system.

Leveraging the unmatched power of APS-ONE-100 compute nodes and the unparalleled KAI IB capabilities, users can address high-performance test scenarios within the most challenging deployment requirements. Teams can achieve complex test topologies by combining APS-ONE-100 compute nodes with virtual test agents, building hybrid test environments or even geographically distributed topologies crossing cloud interconnects.

KAI Inference Builder Specifications

Key Specifications	Options

Deployment options for agents	• Public clouds:
	  ◦ Amazon Web Services (AWS)
	• Prive clouds / on-prem:
	  ◦ VMware ESXi 8,0. 7.0 and ESXi 6.X

Deployment options for controller	• Public clouds:
	  ◦ Amazon Web Services (AWS)
	• Prive clouds / on-prem:
	  ◦ VMware ESXi 8,0. 7.0 and ESXi 6.X

Supported SUT configurations	• Qualified with the following inference engines:
	  ◦ vLLM
	• Supported prompt API format:
	  ◦ OpenAI API – chat completions
	• Qualified with the following LLM models:
	  ◦ GPT-OSS
	  ◦ Llama
	  ◦ Qwen
	  ◦ Mistral

Objective type support	• Prompts per second • Simulated users (secondary)

Prompt type support	• Real-world prompt behavior for:
	  ◦ Financial services
	  ◦ Legal services
	• Inference technology benchmarking
	  ◦ GPU compute

Encryption support	TLS 1.3 with major ciphers and key sizes supported: • AES128-GCM-SHA256 • AES256-GCM-SHA384 • CHACHA20-POLY1305-SHA256 TLS 1.2 with major ciphers and key sizes supported: • AES128-GCM-SHA256 • AES256-GCM-SHA384 • ECDHE-ECDSA-AES128-GCM-SHA256 • ECDHE-ECDSA-AES128-SHA256 • ECDHE-ECDSA-AES256-GCM-SHA384 • ECDHE-ECDSA-AES256-SHA384 • ECDHE-RSA- AES128-GCM-SHA256 • ECDHE-RSA- AES256-GCM-SHA384 • ECDHE-RSA-AES128-SHA256

Key Specifications	Options
	• ECDHE-RSA-AES256-SHA384
Key performance indicators	• Concurrent simulated users • Time to First Token - Max and percentiles (for example, P50, P90, P99) • Time to Last Token - Max and percentiles (for example, P50, P90, P99) • Prompts per Second
Inference & GPU server statistics	• Token rate • Prefill and decode time • Cache utilization • Scheduler state • GPU power usage • Tensor core utilization
Reporting	• PDF • CSV
Max IP addresses per agent	• 10,000
Automation	• Comprehensive REST API coverage • REST API documentation

Product Ordering Information

Part Number	Description
952-1001	KAI Inference Builder Bundle with 2 Agents and up to 100 prompts per second (1-year subscription, floating worldwide). TAA Compliant (952-1001). AI Inference Performance and Security Testing Bundle. The bundle includes: • Access to the KAI Inference Builder cloud-native software application • Up to 2 KAI IB test agents deployable on the customer's environment • Single performance unit count of up to 100 prompts per second and 1000 simulated users: can be consumed in a single concurrent test • Access to ATI, software updates, and customer support for the purchased term of the subscription. REQUIRES: License term to be specified (MUST be purchased in multiples of years). List price is per unit, per year. TAA Compliant.
952-1010	KAI Inference Builder Bundle with 10 Agents and up to 1000 prompts per second (1-year subscription, floating worldwide). TAA Compliant (952-1010). AI Inference Performance and Security Testing Bundle. The bundle includes: • Access to the KAI Inference Builder cloud-native software application • Up to 10 KAI IB test agents deployable on the customer's environment • Single performance unit count of up to 1000 prompts per second and 10,000 concurrent users: can be consumed in a single concurrent test • Access to ATI, software updates, and customer support for the purchased term of the subscription. REQUIRES: License term to be specified (MUST be purchased in multiples of years). List price is per unit, per year. TAA Compliant.
952-1100	KAI Inference Builder Bundle with 10 Agents and up to 10,000 prompts per second (1-year subscription, floating worldwide). TAA Compliant (952-1100). AI Inference Performance and Security Testing Bundle. The bundle includes: • Access to the KAI Inference Builder cloud-native software application • Up to 10 KAI IB test agents deployable on the customer's environment • Single performance unit count of up to 10,000 prompts per second and 100,000 simulated users: can be consumed in a single concurrent test • Access to ATI, software updates, and customer support for the purchased term of the subscription. REQUIRES: License term to be specified (MUST be purchased in multiples of years). List price is per unit, per year. TAA Compliant.