AI-in-a-Box (AiB)
Proof of Concept on Red Hat

QUICKSTART AI PLATFORM

Single node server for customers and partners to explore Gen AI

ROBUST CAPABILITIES

Out-of-box capabilities to perform:

- LLM Model customization
- LLM Model Inferencing
- Simple LLM Demo Application

EXTENSIVELY TESTED

Rigorously tested against widely-used industry & enterprise-grade Large Language Models

JOINT COLLABORATIONS

Validated by:

- Cloud Native Architecture (CNA)
  Email: ask.cna@dell.com

Supported by:

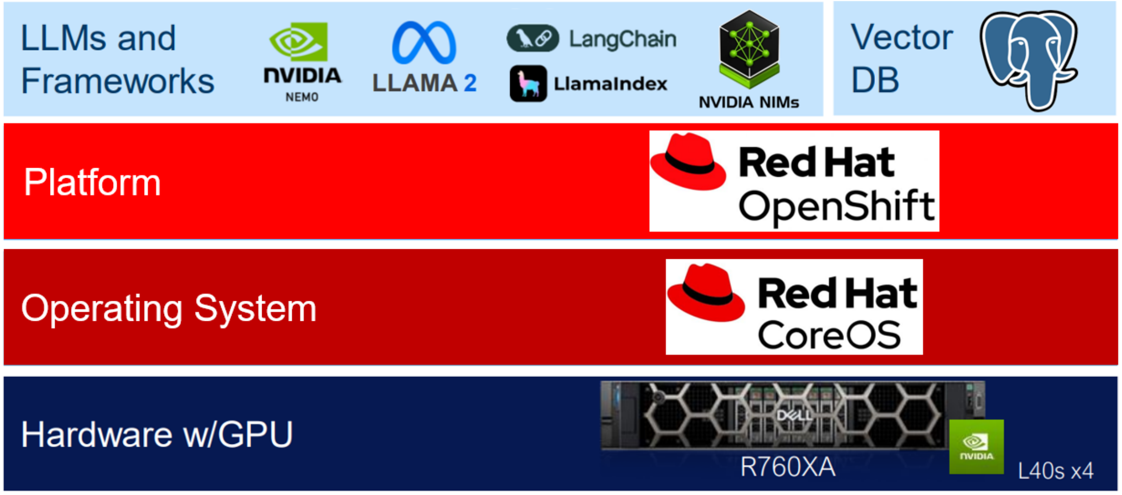
- CSC Singapore
- Global Alliance
- DS@DCWS APJ

Proof-of-Concept

1.1 Overview of AI-in-a-Box(AiB)

AI-in-a-Box is a quick-start single-node server option for customers and partners to explore Generative AI workloads and use cases. It is validated to provide sufficient performance for model training and inferencing tasks required for proof of concept use cases.

1.2 High-Level Solution Architecture



1.3 Solution Components

Hardware	
Server	PowerEdge R760xa
CPU	Intel(R) Xeon(R) Gold 6448Y x 2 (32core, 2.1 GHz Base)
Memory	16x 32GB DDR-5 DIMM 4800 MT/s
GPU	4x NVIDIA L40S
Local Storage	OS Drive – 447GB Data Drive – 14 TB (NVMe RAID Disk)
Software	
Platform	OpenShift v4.14 (Single-Node-OpenShift)
Nemo Container	nvcr.io/ea-bignlp/ga-participants/nemofw-training:23.08.03
TensorRT-LLM	v0.7.1
RAG Chatbot	See Section 4.

2. Model Customization

Model customization uses finetuning techniques with domain-specific datasets on pre-trained models to enable domain-specific tasks. In validation tests with Nvidia’s Nemo, AiB's performance is benchmarked by finetuning the Llama 2 model with Databricks' dolly-15k dataset containing 15,000 rows of data.

2.1 Validation Configurations

Each benchmarked model undergoes popular finetuning techniques like Supervised Fine-Tuning (SFT), P-Tuning, and Low Rank Adaptation (LoRA), with performance evaluated based on training time.

	Llama 2 (7B)	Llama 2 (13B)
SFT	Number of GPUs: 4 TP: 4 PP: 1 Maximum no. of steps: 1000	N/A
P-Tuning	Number of GPUs: 2, 4 TP: 2, 4 PP: 1 Maximum no. of steps: 1000	Number of GPUs: 4 TP: 4 PP: 1 Maximum no. of steps: 1000
LoRA	Number of GPUs: 2, 4 TP: 1 PP: 2, 4 Maximum no. of steps: 1000	N/A

TP – Tensor Parallelism
PP – Pipeline Parallelism

2.2 Validation Results

Model	No. of GPUs	SFT	P-Tuning	LoRA
Llama 2 (7B)	2	**N/A**	678	447
	4	642	563	248
Llama 2 (13B)	4	**N/A**	919	**N/A**

Values are time to fine-tune model in ***minutes***
Note: These timings exclude the loading of the mode, the dataset and model validation.

ROBUSTLY VALIDATED

Rigourously tested for model customization with widely-used industry & enterprise grade Large Language Models and customization techniques

EXTENSIVE MODEL CUSTOMIZABILITY

Widely applicable to common industry model fine-tuning standards & methods:

- Supervised Fine-Tuning (SFT)
- P-Tuning
- Low Rank Adapters (LoRA)

ALL-ROUND CAPABILITY

Supports GenAI Inferencing on myriad of model sizes, up to large-sized models **of 70B parameters**

THROUGHOULY TESTED

Extensively tested across various production inferencing use cases, capable of utilizing high-end LLMs in multiple scenarios.

USE CASES EXAMPLES

- Intelligent documentation creation and processing
- Code generation, assistance, and documentation
- Content creation or chatbots for marketing and sales
- Sentiment analysis
- Virtual Customer service agents

Proof-of-Concept

3. Model Inferencing

In the validation tests, the performance of AiB is benchmarked using NVIDIA NeMo Frameworks with NVIDIA TensorRT-LLM. Latency is then measured on the Llama2 models with inputs of varying token lengths (128 and 2048).

3.1 Observed GPU Memory Consumptions

Model	Quantization	GPU Memory Consumption (GB)
Llama 2 (7B)	FP8	24.3
	AWQ	16.6
Llama 2 (13B)	FP8	34.9
	AWQ	23.0
Llama 2 (70B)	FP8*	133.9
	AWQ**	70.3

Note: Measured with a batch size of 1, input length of 128 and output length of 1

* Measured based on 4x L40S

** Measured based on 2x L40S

3.2 Validation Results (First Token Latency)

Model	Quantization	1x L40S		2x L40S		4x L40S	
		Token Length					
		128	2048	128	2048	128	2048
Llama 2 (7B)	FP8	19.7	93.7	17.1	122.3	16.4	123.4
	AWQ	14.7	158.5	15.6	148.5	**N/A**	**N/A**
Llama 2 (13B)	FP8	31.8	178.8	27.8	206.0	24.4	202.1
	AWQ	26.1	329.6	24.6	266.3	21.3	230.6
Llama 2 (70B)	FP8	**N/A**	**N/A**	**N/A**	**N/A**	79.3	696.2
	AWQ			92.7	1091.7	68.4	847.5

Values are time recorded in ***milliseconds***

3.3 Validation Results (Throughput)

Model	Quantization	1x L40S	2x L40S	4x L40S
Llama 2 (7B)	FP8	2966.9	3617.9	4298.8
	AWQ	2900.8	3575.3	**N/A**
Llama 2 (13B)	FP8	1650.7	2140.9	2616.3
	AWQ	1746.1	2215.1	2651.0
Llama 2 (70B)	FP8	**N/A**	**N/A**	787.1
	AWQ		712.9	949.3

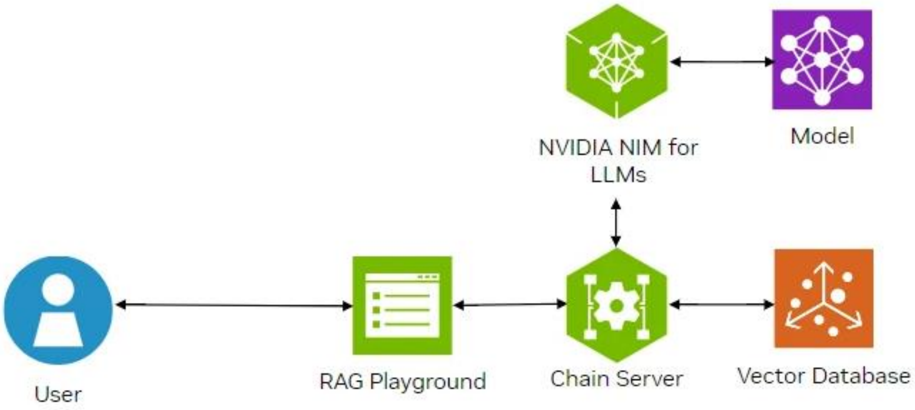
Values represents ***tokens per second***

Note: Measured with a batch size of 64, input length of 128 and output length of 128

4. Sample Use Case: Q&A Chatbot

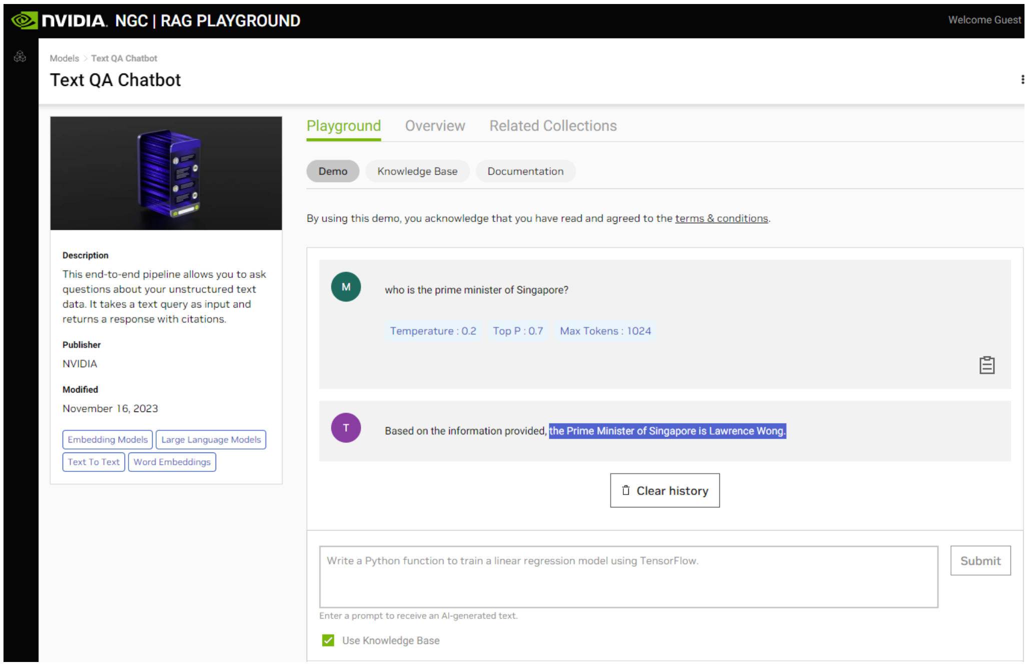
A RAG chatbot application has been deployed on AiB as part of the PoC using the NVIDIA RAG LLM Operator. The chatbot is powered by the Llama2 13b-chat model running on NIMs.

4.1 Software Components



Component	Detail
Inference Server	Triton Inference Server
LLM Engine	vLLM
Pretrained LLM Model	Llama-2 13B-chat
Embedding Model	NV-Embed-QA-4
Vector Database	pgvector
User Interface	NVIDIA RAG Playground
Data Frameworks	LlamaIndex, LangChain
RAG Pipeline	RAG LLM Operator with NIMs

4.2 RAG ChatBot



Description

This end-to-end pipeline allows you to ask questions about your unstructured text data. It takes a text query as input and returns a response with citations.

Publisher

NVIDIA

Modified

November 16, 2023

By using this demo, you acknowledge that you have read and agreed to the terms & conditions.

M: who is the prime minister of Singapore?

Temperature : 0.2 Top P : 0.7 Max Tokens : 1024

T: Based on the information provided, the Prime Minister of Singapore is Lawrence Wong.

Write a Python function to train a linear regression model using TensorFlow.

Enter a prompt to receive an AI-generated text.

☑ Use Knowledge Base

VERSATILITY ACROSS INDUSTRY USE CASES

AI-in-a-Box handles end-to-end GenAI tasks, spanning various industry applications. From basic chatbots using pre-trained models to tailored RAG pipelines with custom models.

EASE OF DEPLOYMENT

Leverage the NVIDIA RAG LLM Operator to simplify the deployment of production-ready RAG pipelines on RedHat OpenShift.

Proof-of-Concept

5. Power Consumption

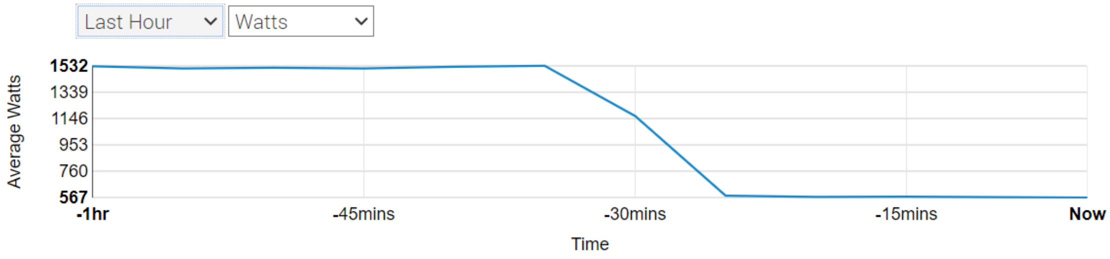
The AiB is a product configured and designed with sustainability and efficiency at its core. This cutting-edge, all-rounded, quick-starter package not only brings the power of artificial intelligence to your fingertips but also does so while adhering to most data center power consumption regulations and policies.

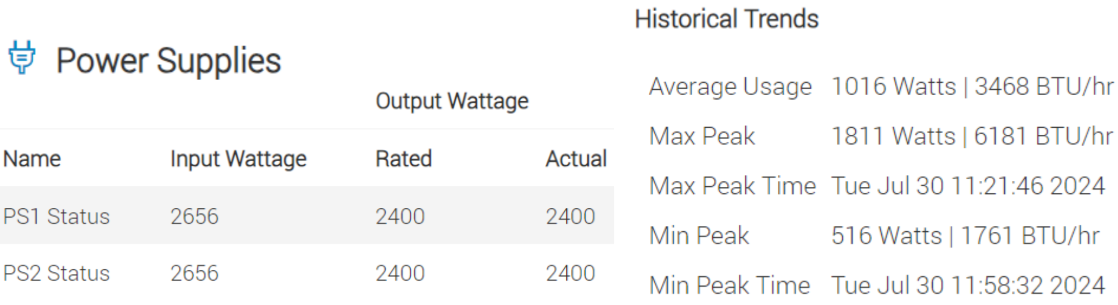
With a maximum peak power consumption load of **1811 watts** or **6181 BTU/hr**, the AI-in-a-box ensures optimal performance without compromising on energy efficiency. Harness the power of AI while staying green and most importantly compliant with your current existing environment, without needless drastic environment changes.

5.1 Peak Training Power Consumption

To ensure and assure the AiB is capable of handling most demanding recommended workloads without making drastic datacenter changes to accommodate, the AiB has been rigorously tested under intensive workloads and monitored throughout.

Tests involved fine-tuning the Large Language Model (LLM) utilizing all four GPUs to their maximum capacity. Throughout these tests, power consumption metrics were continuously monitored, ensuring that even under the most demanding conditions, the AiB stays within its specified power consumption limits, ultimately compliant to common datacenter power limits.





		Output Wattage	
Name	Input Wattage	Rated	Actual
PS1 Status	2656	2400	2400
PS2 Status	2656	2400	2400

Average Usage	1016 Watts | 3468 BTU/hr
Max Peak	1811 Watts | 6181 BTU/hr
Max Peak Time	Tue Jul 30 11:21:46 2024
Min Peak	516 Watts | 1761 BTU/hr
Min Peak Time	Tue Jul 30 11:58:32 2024

6. Conclusion

Dell’s AiB offers a compact yet robust single-node server for customers or partners delving into GenAI solutions. Satisfactory test results on model customization and inferencing affirm that AiB’s performance remains uncompromised by its size, fully supporting LLM workloads.

Moreover, the AiB, along with its tested power consumption, forms a convenient package that is compliant with current data center standards. This ensures that Dell’s AiB can be seamlessly integrated into existing infrastructure, providing a practical and efficient solution for those seeking to leverage the power of GenAI. At the same time, Dell’s AiB continues to uphold its commitment to sustainability and energy efficiency, further enhancing its appeal to modern data centers.

DEPLOYABLE OUT-OF-BOX

Tuned to maximize GPU compute, ensuring optimal performance and power efficiency under demanding LLM workloads, all while fitting seamlessly into your existing data center environment without any changes.

STAY COMPLIANT WITHIN POWER CONSUMPTION

Ensuring compliance with common datacenter power limits, even under the most demanding conditions.

Proof-of-Concept