OPEA-based Retrieval Augmented Generation (RAG) on Intel® Gaudi with OpenShift AI

Red Hat Summit Connect 2024 Denmark

Copenhagen, 29 October 2024

Red Hat | intel.

Red Hat

Codrin Bucur

Principal AI Specialist Solution Architect,
EMEA, Red Hat

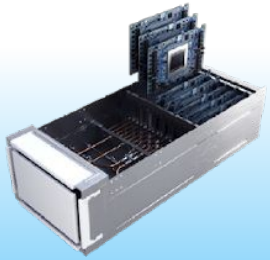
Red Hat + intel.

Over 25 Years of Collaboration

Bringing AI Everywhere

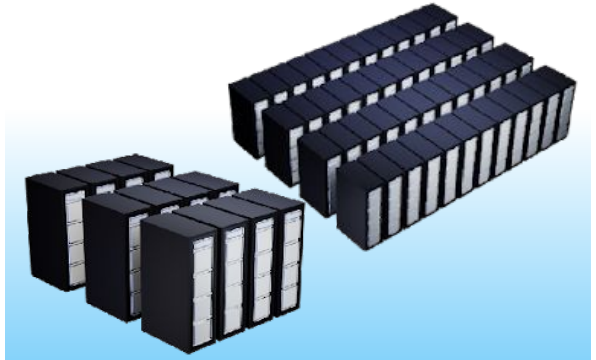
Intel's AI Strategy

AI PC Node
AI Developer Productivity & Light Inference



Node
Fine-tuning, Inference

Cluster
Light Training, Tuning, Peak Inference



Super Cluster
Training, Tuning, Peak Inference

Mega Cluster
Large Scale Training & Inference

AI PC
Broadest AI SW Ecosystem

ENTERPRISE AI & EDGE AI
Open Standard, "Ready to Use"

DATA CENTER AI
AI Open, Scalable Systems & Reference Arch

intel CORE ULTRA | intel ARC GRAPHICS | intel CORE ULTRA | intel XEON | intel GAUDI | intel XEON | intel GAUDI | intel IPU

Red Hat | intel

Red Hat's AI Strategy

Trust | **Choice** | **Consistency**

AI models

RHEL AI

Base Model | Alignment Tuning | Methodology & Tools | Platform Optimization & Acceleration

AI platform

OpenShift AI

Development | Serving | Monitoring & Lifecycle | MLOps | Resource Management

AI enabled portfolio

Lightspeed portfolio

Usability & Adoption | Guidance | Virtual Assistant | Code Generation

AI workload support

Optimize AI workloads

Deployment & Run | Compliance | Certification | Models | Open Source Ecosystem

Open Hybrid Cloud Platforms

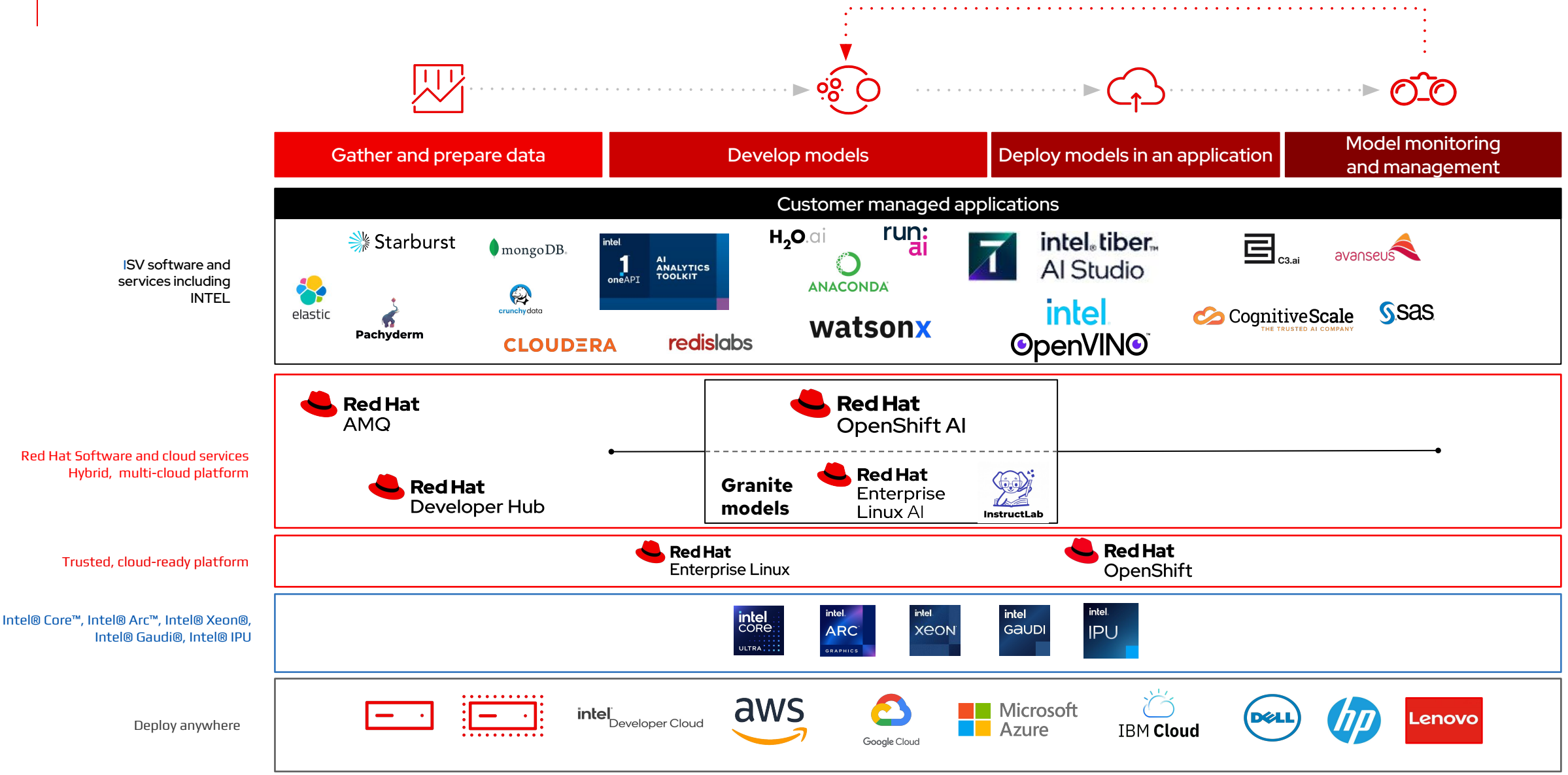
Red Hat Enterprise Linux | Red Hat OpenShift | Red Hat Ansible Platform

Acceleration | Performance | Scale | Automation | Observability | Security | Developer Productivity | App Connectivity | Secure Supply Chain

Partner Ecosystem

Hardware | Accelerators | Delivery

Intel Enterprise AI with Red Hat® OpenShift® AI



Red Hat | intel.

OPEA – Open Platform for Enterprise AI

Red Hat | intel.

OPEA - Open Platform for Enterprise AI

By The Linux Foundation

- Ecosystem orchestration framework for GenAI
- OPEA.dev
- GitHub: https://github.com/opea-project
- Contributors:

OPEA with OpenShift AI

OpenShift AI makes OPEA more enterprise ready

Open Platform for Enterprise AI

RAG Workflow

Framework building blocks

Red Hat OpenShift

Pipelines
Service Mesh
Telemetry

Red Hat OpenShift AI

Advanced AI Tools
Model Serving

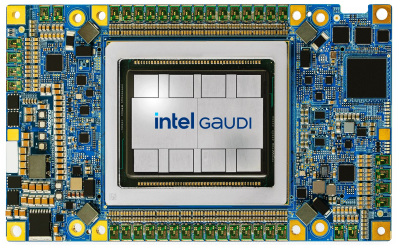
intel XEON & intel GAUDI

Advanced AI Acceleration

Intel Gaudi
AI Accelerators

Red Hat | intel.

Introducing the Intel® Gaudi® 3 Accelerator



Breaking benchmarks, not budgets

Competitive Gen AI Performance over H100

- Projected **50% faster time to train**[1]
- Projected **50% faster inferencing**[2]
- Projected **40% better power efficiency**[3]

Freedom to Scale without Lock-in

- Open standard ethernet networking vs proprietary InfiniBand
- 24x200 GbE ports of industry-standard RoCE on every Gaudi® [3]
- 33% more I/O peak throughput vs H100 for massive scale-up within the server[4]

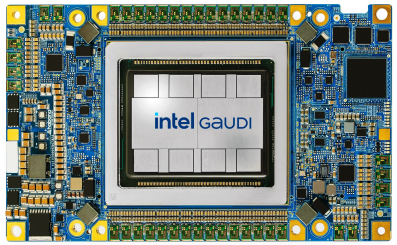
Open Development on GenAI platforms

- Integrated open-source PyTorch framework with optimized model library on Hugging Face
- Migrate models on open software from H100 with as few as 3 lines of code

[1] NV H100 comparison based on : https://developer.nvidia.com/deep-learning-performance-training-inference/training, Mar 28th 2024 -> “Large Language Model” tab.
[2] Source: NV H100 comparison based on https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8 , Mar 28th, 2024. Reported numbers are per GPU.
[3] Source: NV comparison based on https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8 , Mar 28th, 2024. Reported numbers are per GPU.
[1-3] Vs Intel® Gaudi® 3 projections for LLAMA2-7B, LLAMA2-70B & Falcon 180B Power efficiency for both Nvidia and Gaudi3 based on internal estimates. Results may vary.
[4] 900 GB/s NVLink connectivity on H100 vs. 1200 GB/s on Gaudi 3

Intel Gaudi AI Accelerators

Broad Application Support with Focus on Multi-Modal, LLM and RAG



AI Applications

AI Functions

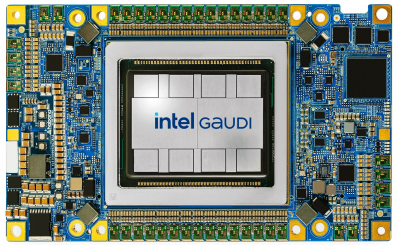
3D Generation	Text Generation	Classification
Video Generation	Sentiment	Translation
Image Generation	Summarization	Q&A

Core Capabilities

- Multi-modal Models
- LLM
- RAG

Intel® Gaudi® 3 AI Accelerator

Launch Partners



DELL Technologies

Lenovo

Hewlett Packard Enterprise

SUPERMICRO

ASUS

Inventec

QCT Quanta CLOUD TECHNOLOGY

GIGABYTE TECHNOLOGY

ingrasys

wistron

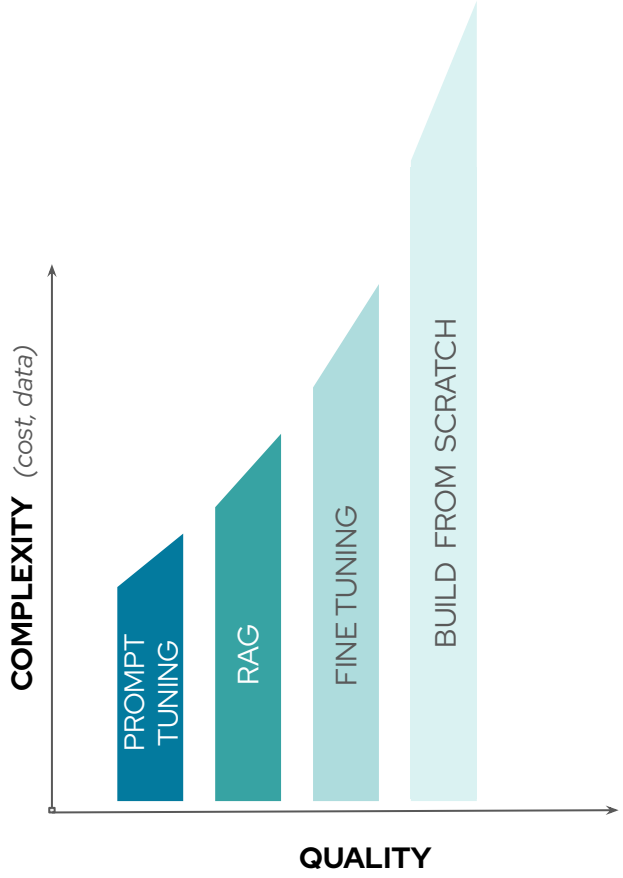
Retrieval Augmented Generation (RAG) Explained

Red Hat | intel.

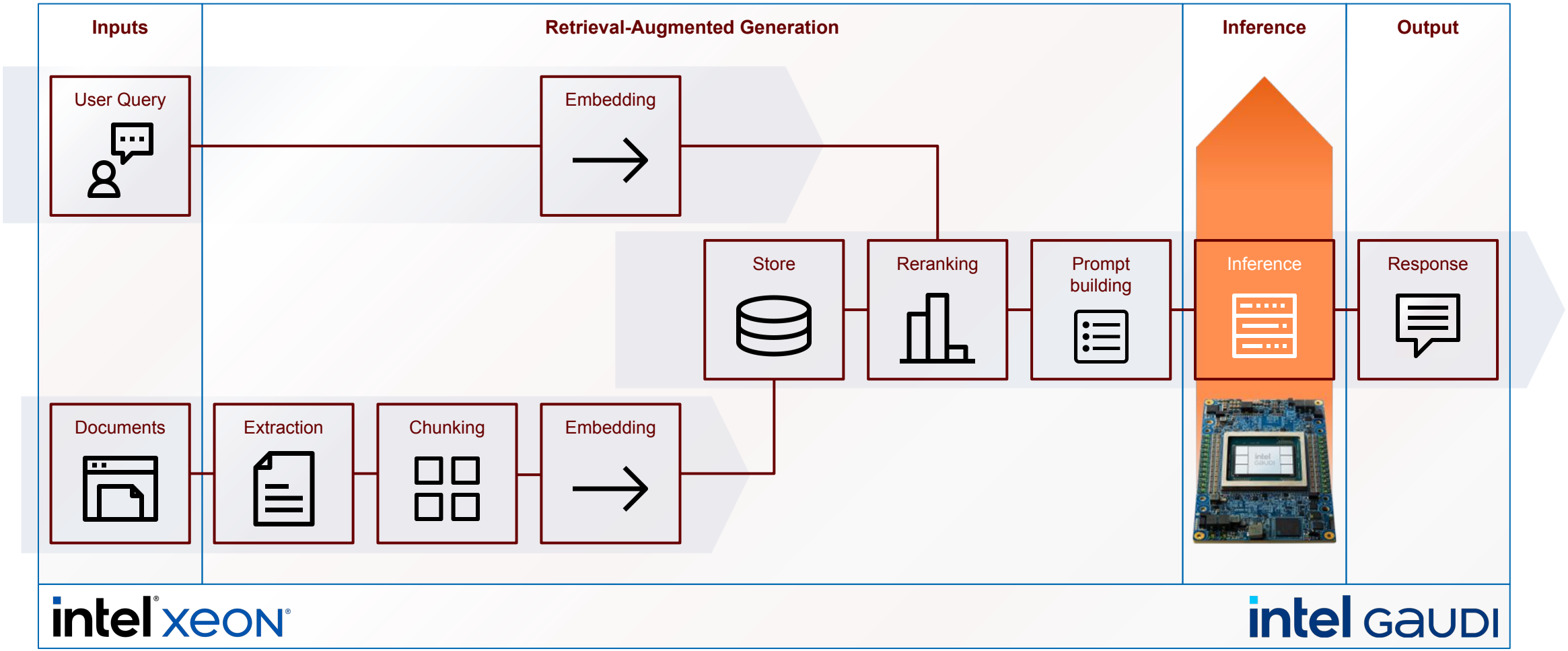
The balancing act of using foundation models

Foundation models will still need more work to be useful

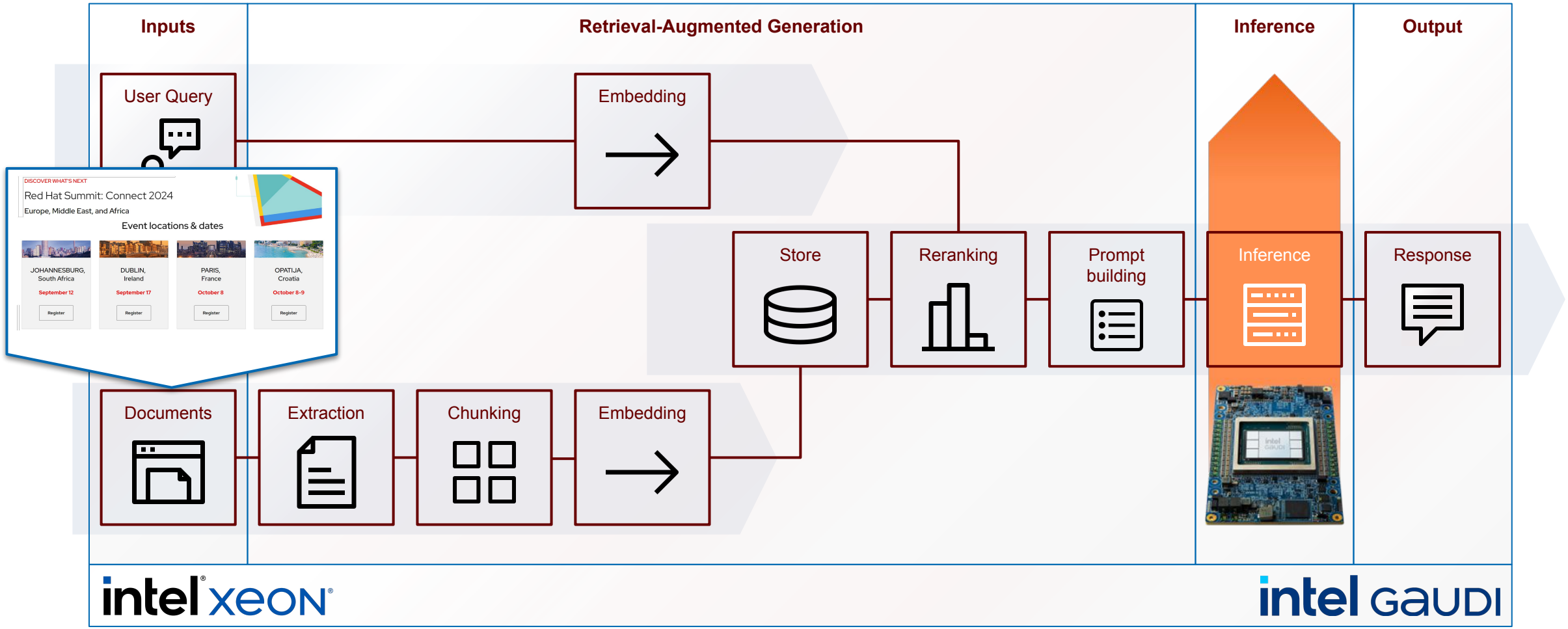
- Prompt tuning
- Retrieval-Augmented Generation (RAG)
- Fine tuning foundation models
- Training a Foundation Model from scratch



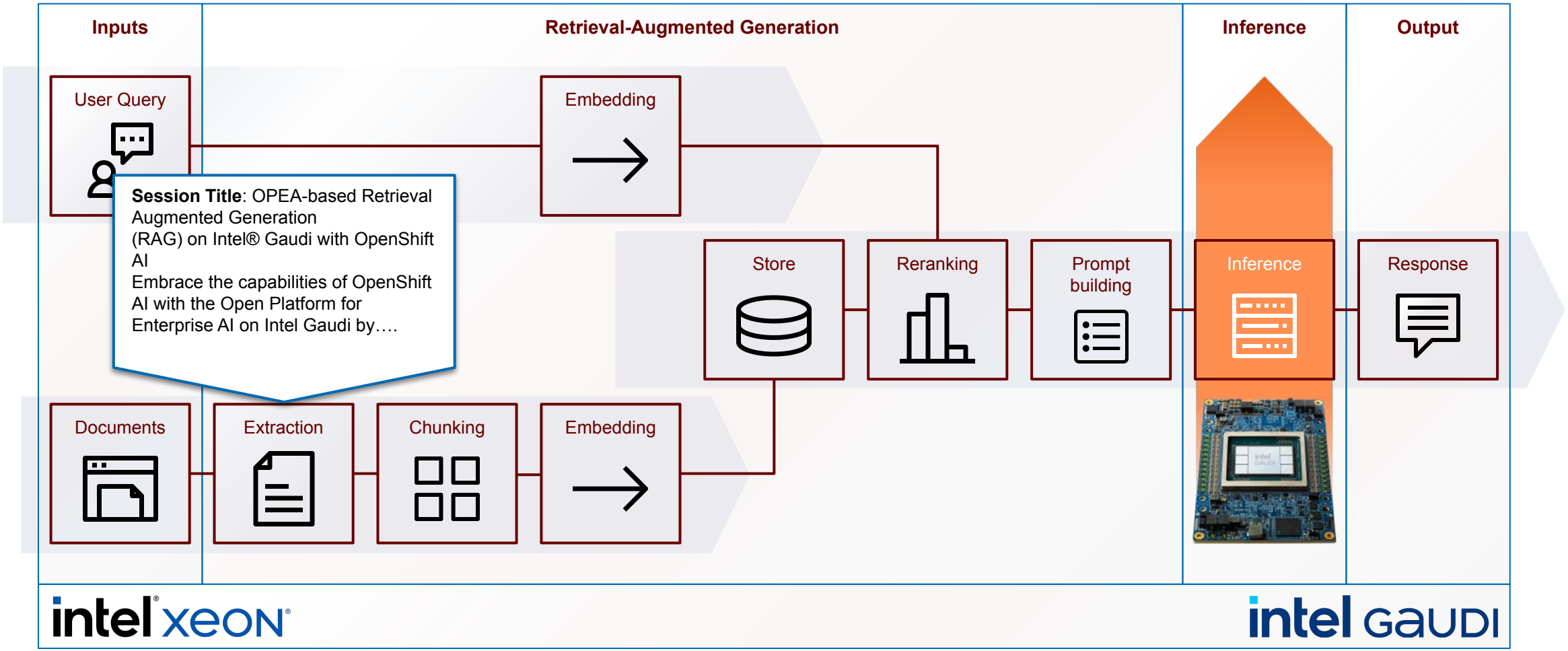
Retrieval Augmented Generation (RAG)



Retrieval Augmented Generation (RAG)

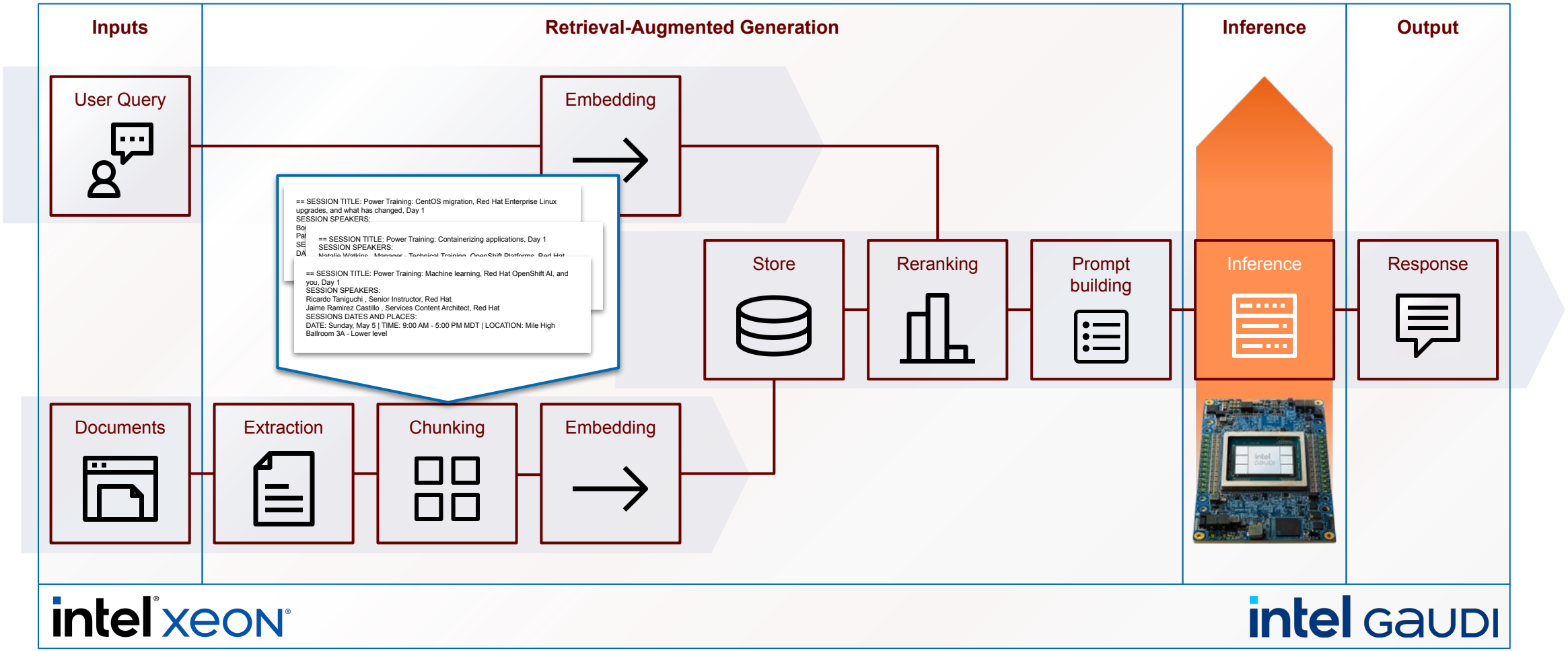


Retrieval Augmented Generation (RAG)

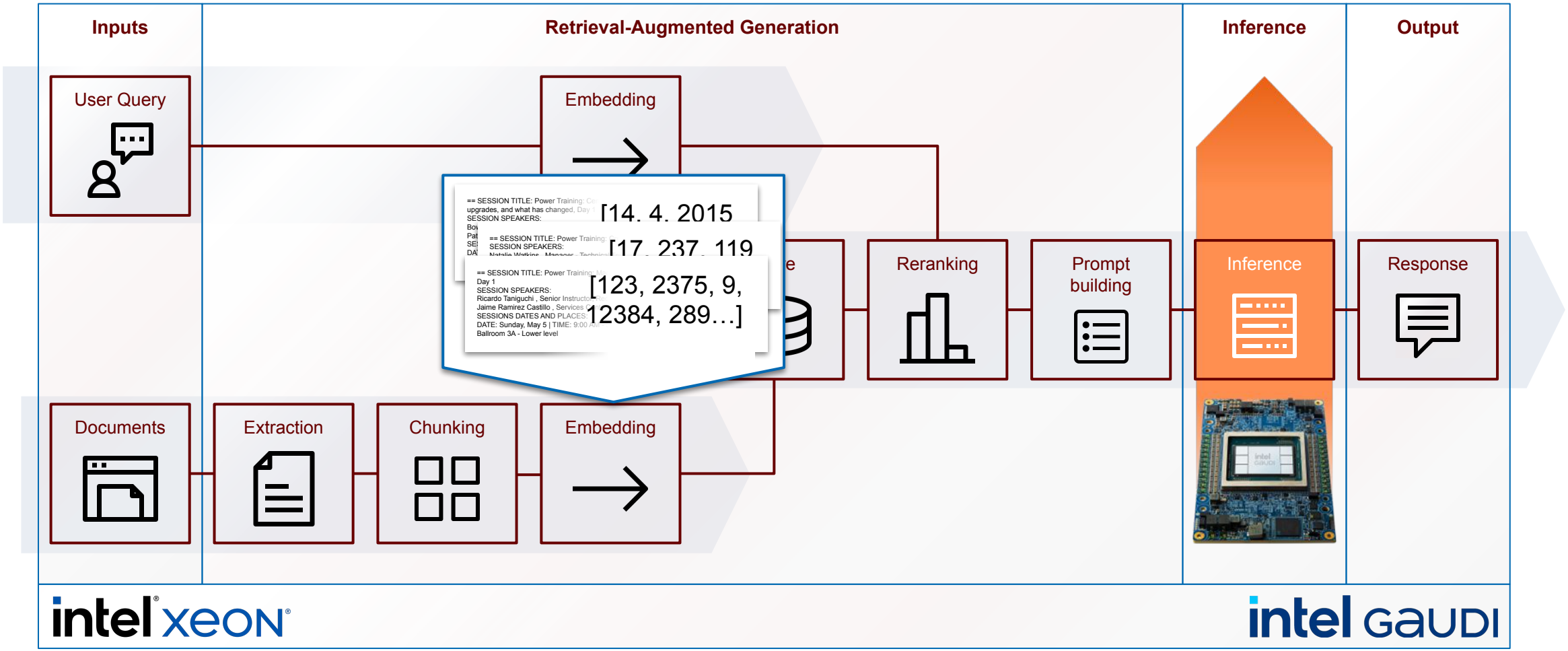


Session Title: OPEA-based Retrieval Augmented Generation (RAG) on Intel® Gaudi with OpenShift AI
Embrace the capabilities of OpenShift AI with the Open Platform for Enterprise AI on Intel Gaudi by….

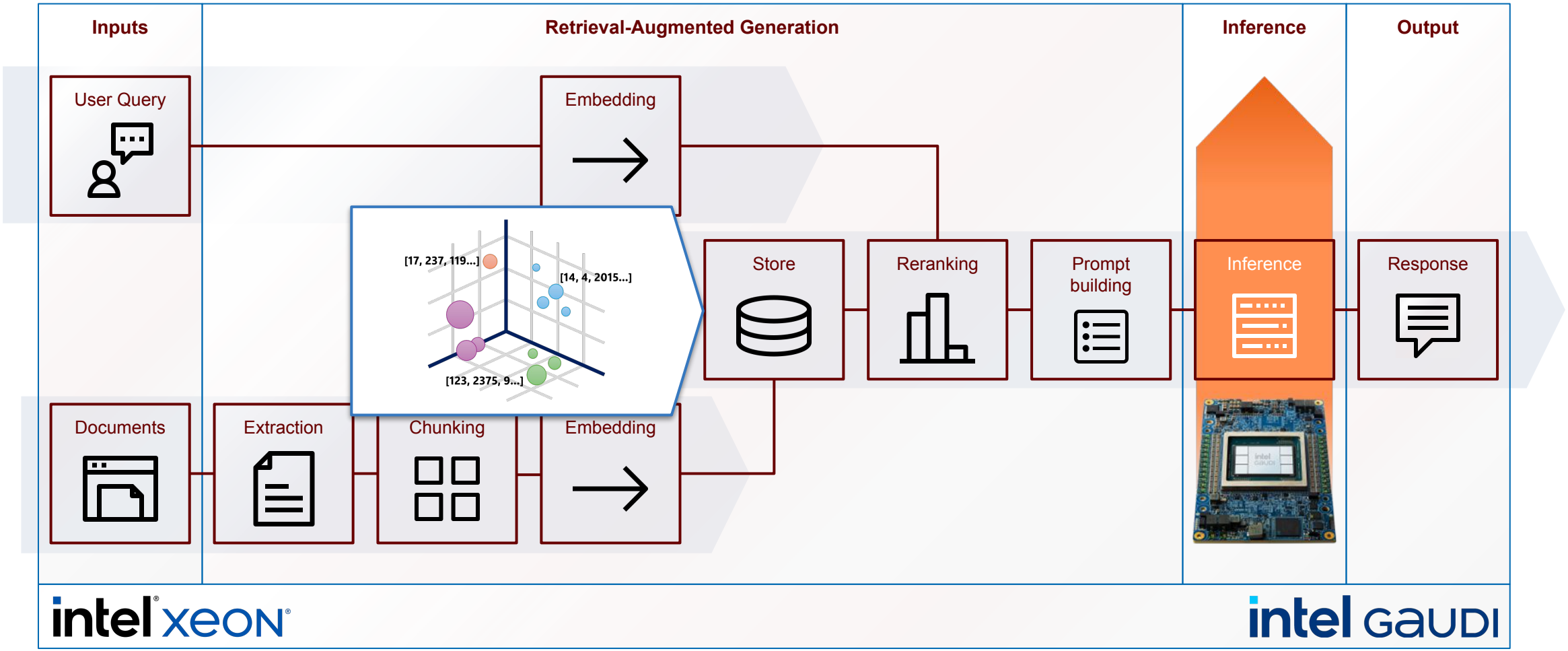
Retrieval Augmented Generation (RAG)



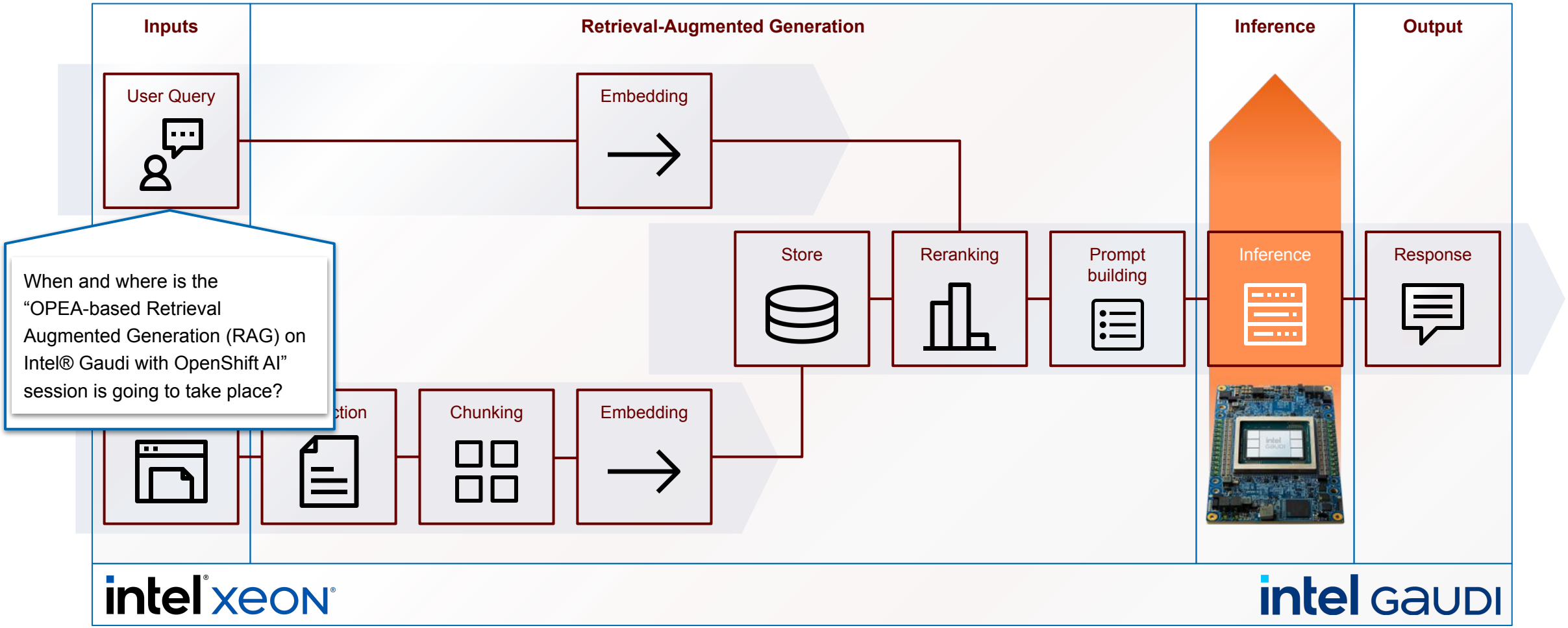
Retrieval Augmented Generation (RAG)



Retrieval Augmented Generation (RAG)

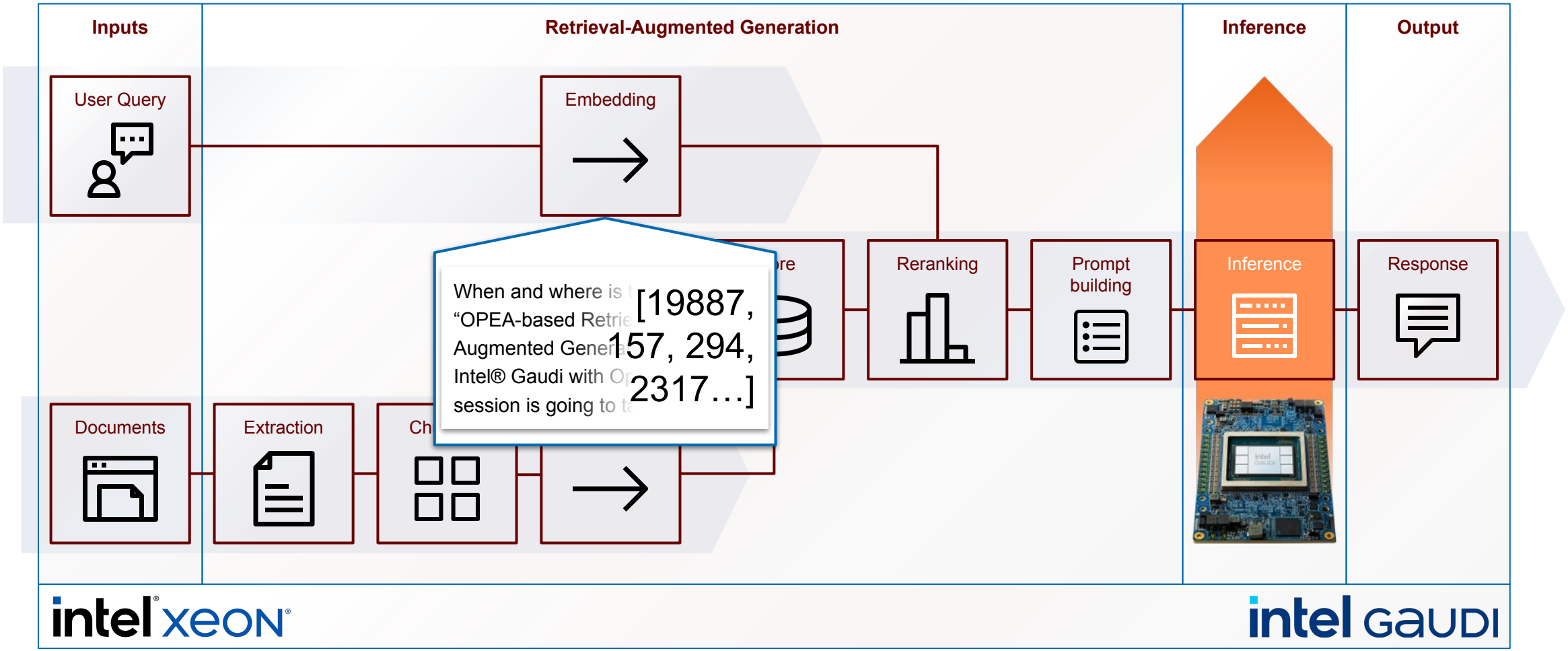


Retrieval Augmented Generation (RAG)

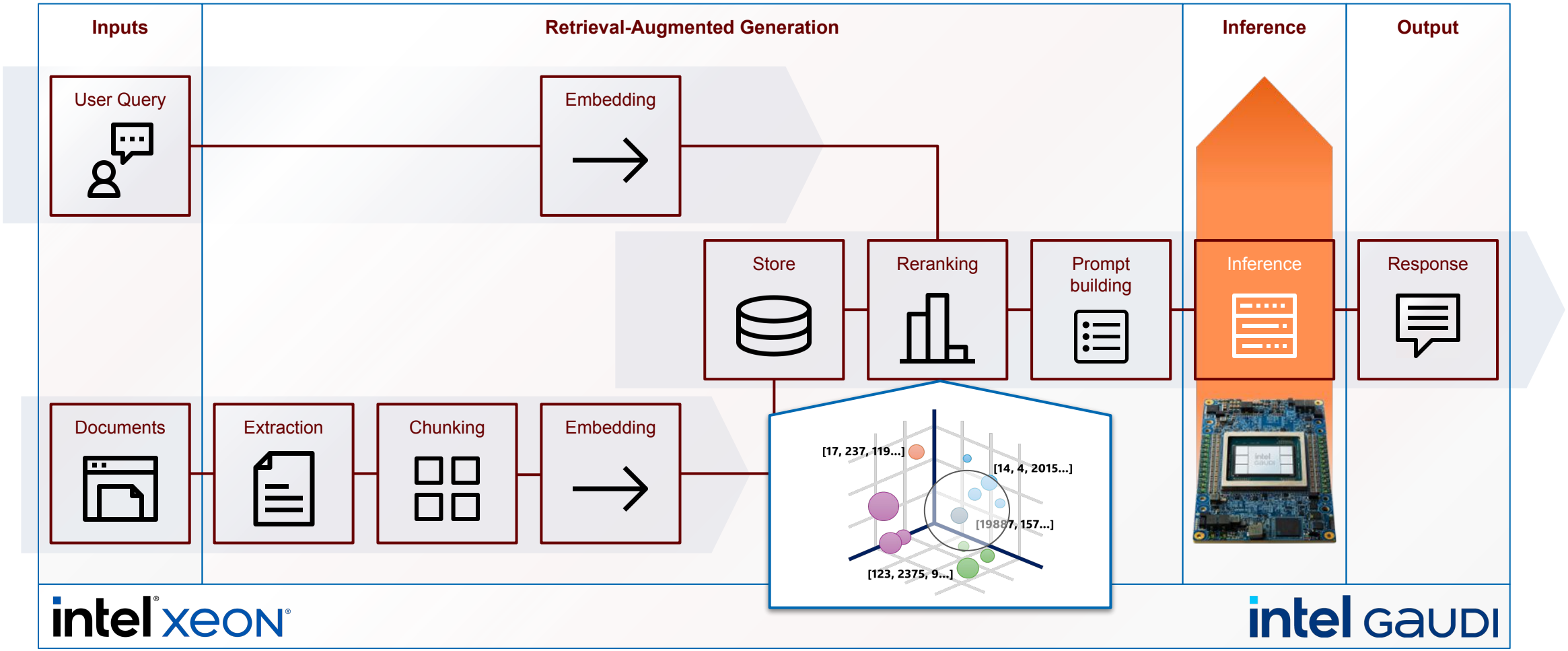


When and where is the "OPEA-based Retrieval Augmented Generation (RAG) on Intel® Gaudi with OpenShift AI" session is going to take place?

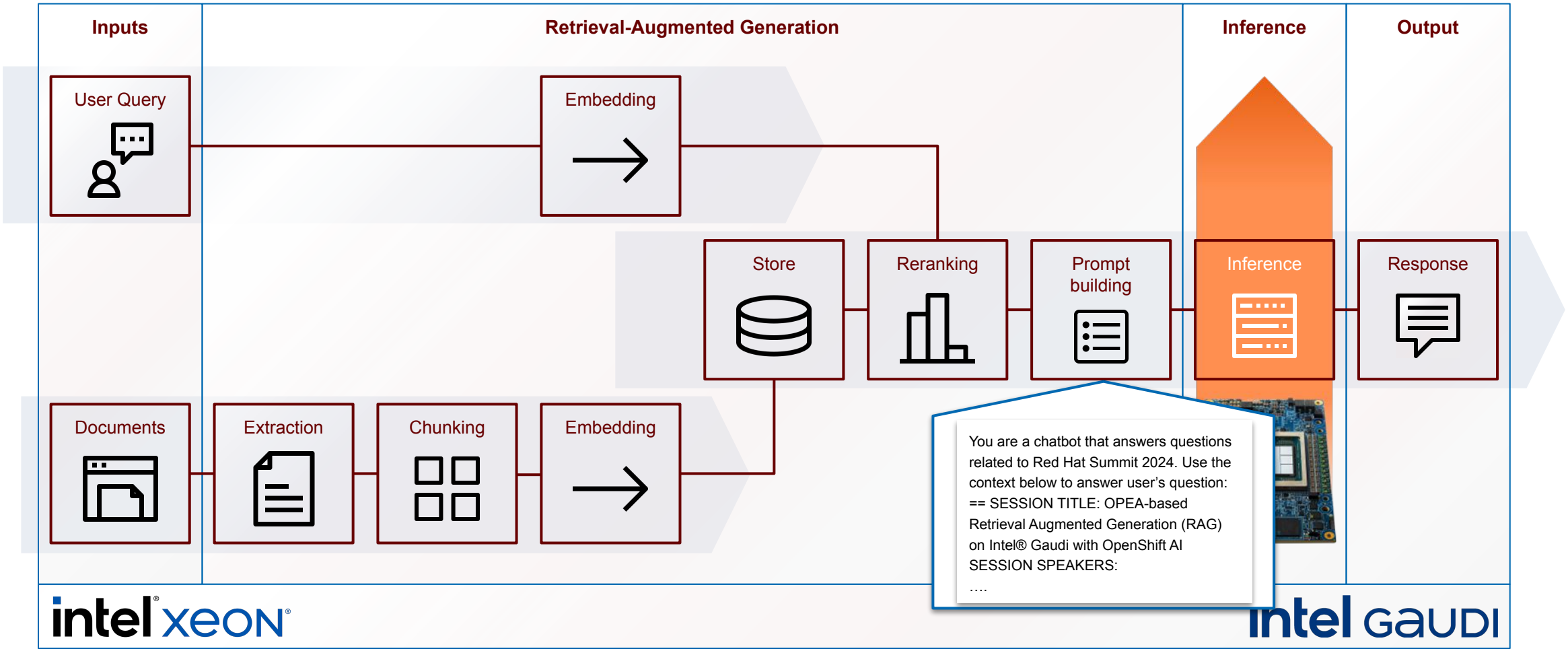
Retrieval Augmented Generation (RAG)



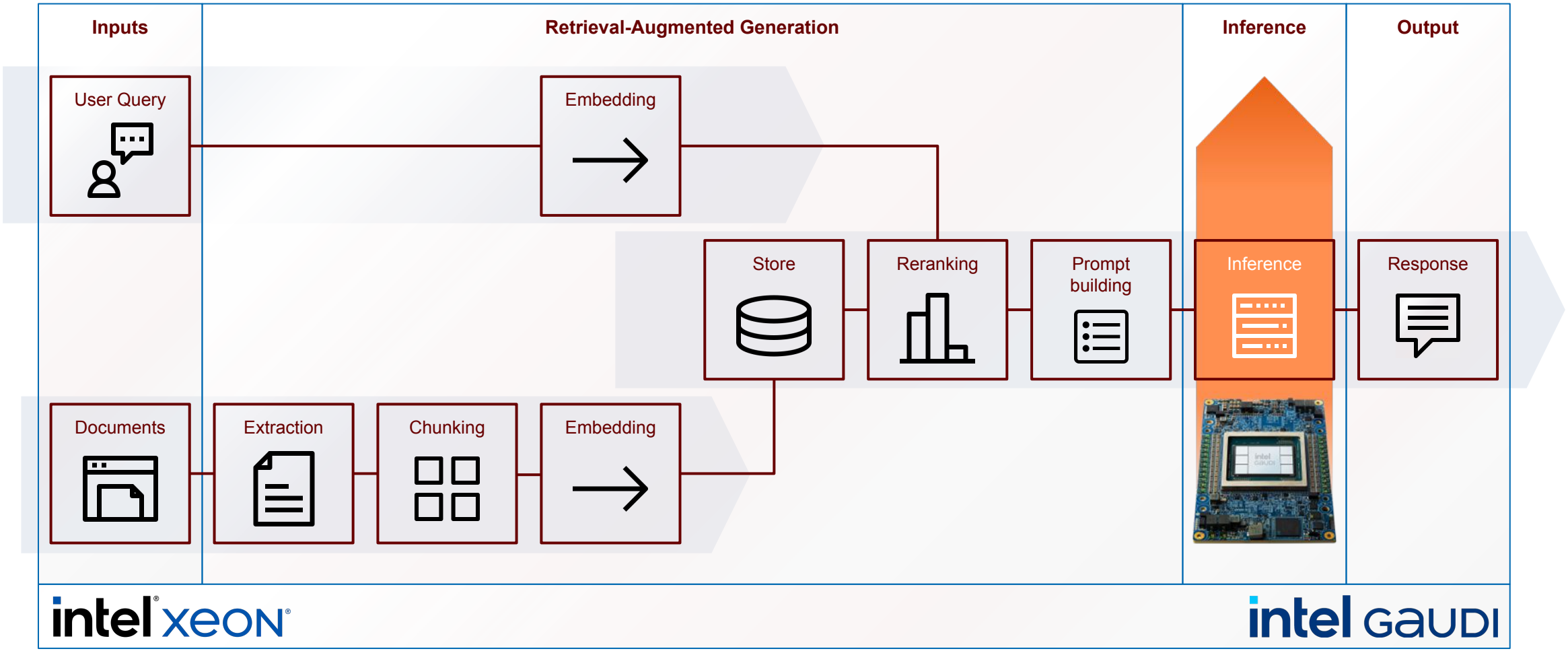
Retrieval Augmented Generation (RAG)



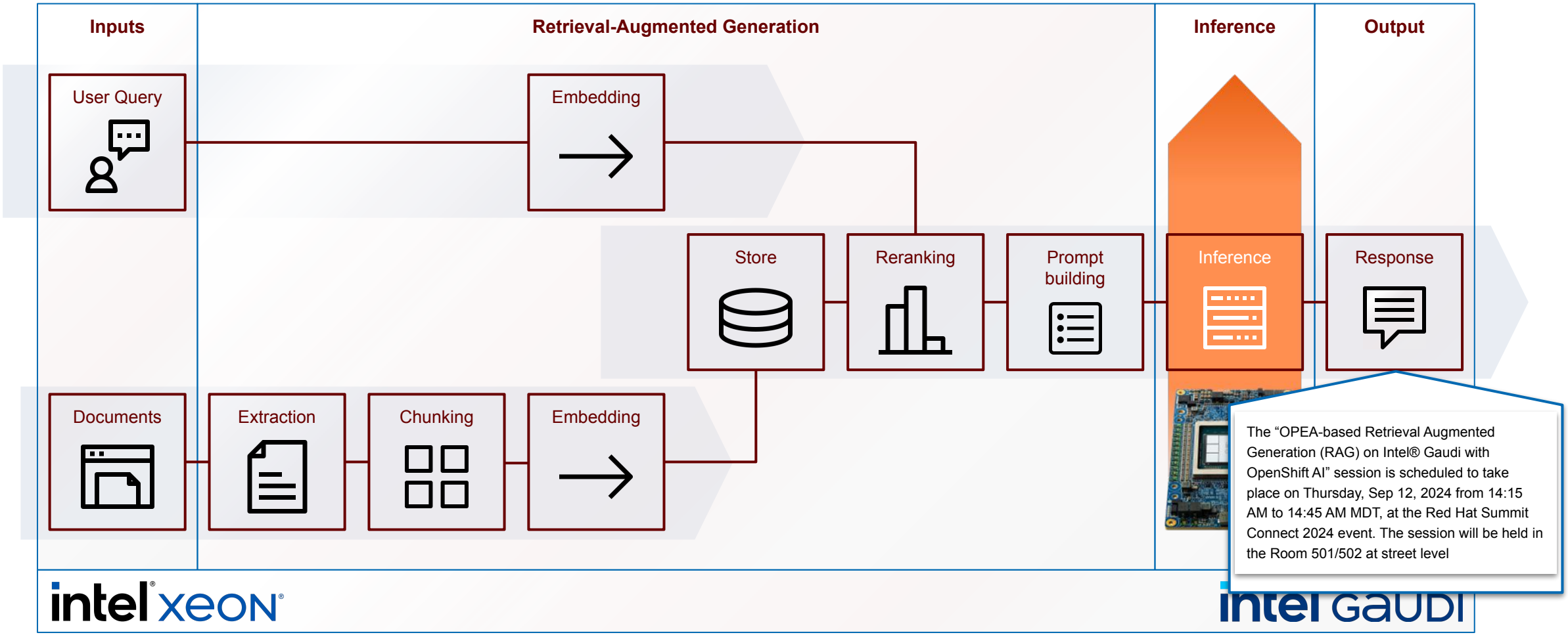
Retrieval Augmented Generation (RAG)



Retrieval Augmented Generation (RAG)



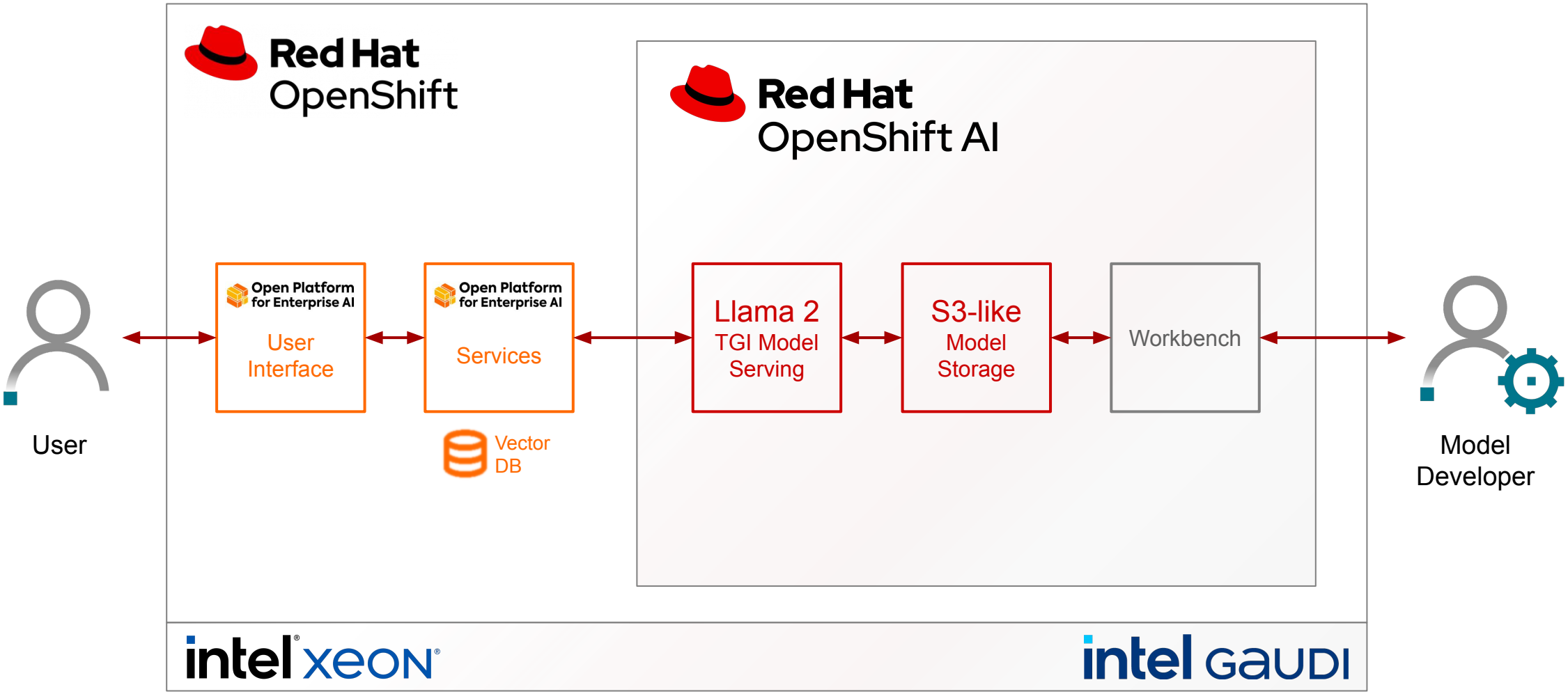
Retrieval Augmented Generation (RAG)



Retrieval Augmented Generation (RAG) Chatbot Demo

Red Hat | intel.

OPEA Chat QnA on OpenShift – demo design



Red Hat OpenShift

5 kube:admin

Administrator

Home

Operators

OperatorHub

Installed Operators

Workloads

Serverless

Networking

Storage

Builds

Observe

Compute

User Management

Administration

Project: All Projects

Installed Operators

Name Search by name... /

Name	Namespace	Managed Namespaces	Status	Last updated	Provided APIs
Habana AI 1.15.0-479 provided by Habana Labs Ltd.	NS habana-ai-operator	NS habana-ai-operator	Succeeded Up to date	Apr 30, 2024, 6:10 PM	Device Config
Kernel Module Management 2.1.0 provided by Red Hat	NS openshift-kmm	All Namespaces	Succeeded Up to date	Apr 30, 2024, 11:54 AM	PreflightValidation PreflightValidationOCP Module NodeModulesConfig
LVM Storage 4.14.4 provided by Red Hat	NS openshift-storage	NS openshift-storage	Succeeded Up to date	Apr 29, 2024, 3:17 PM	LVMCluster
Node Feature Discovery Operator 4.14.0-202404161544 provided by Red Hat	NS openshift-nfd	NS openshift-nfd	Succeeded Up to date	Apr 30, 2024, 6:10 PM	NodeFeatureDiscovery NodeFeatureRule
Package Server 0.0.1-snapshot provided by Red Hat	NS openshift-operator-lifecycle-manager	NS openshift-operator-lifecycle-manager	Succeeded	Apr 29, 2024, 3:17 PM	PackageManifest
Red Hat OpenShift AI 2.8.1 provided by Red Hat	NS redhat-ods-operator	All Namespaces	Succeeded Up to date	Apr 29, 2024, 3:17 PM	Data Science Cluster DSC Initialization FeatureTracker
Red Hat OpenShift Serverless 1.32.1 provided by Red Hat	NS openshift-serverless	All Namespaces	Succeeded Up to date	Apr 29, 2024, 3:18 PM	Knative Serving Knative Eventing Knative Kafka
Red Hat OpenShift Service Mesh 2.5.1-0 provided by Red Hat, Inc.	NS openshift-operators	All Namespaces	Succeeded Up to date	Apr 30, 2024, 11:54 AM	Istio Service Mesh Control Plane Istio Service Mesh Member Istio Service Mesh Member Roll

Red Hat OpenShift

5 kube:admin

Administrator

Home

Operators

OperatorHub

Installed Operators

Workloads

Serverless

Networking

Storage

Builds

Observe

Compute

User Management

Administration

Project: All Projects

Installed Operators

Name Search by name... /

Name	Namespace	Managed Namespaces	Status	Last updated	Provided APIs
Habana AI 1.15.0-479 provided by Habana Labs Ltd.	NS habana-ai-operator	NS habana-ai-operator	Succeeded Up to date	Apr 30, 2024, 6:10 PM	Device Config
Kernel Module Management 2.1.0 provided by Red Hat	NS openshift-kmm	All Namespaces	Succeeded Up to date	Apr 30, 2024, 11:54 AM	PreflightValidation PreflightValidationOCP Module NodeModulesConfig
LVM Storage 4.14.4 provided by Red Hat	NS openshift-storage	NS openshift-storage	Succeeded Up to date	Apr 29, 2024, 3:17 PM	LVMCluster
Node Feature Discovery Operator 4.14.0-202404161544 provided by Red Hat	NS openshift-nfd	NS openshift-nfd	Succeeded Up to date	Apr 30, 2024, 6:10 PM	NodeFeatureDiscovery NodeFeatureRule
Package Server 0.0.1-snapshot provided by Red Hat	NS openshift-operator-lifecycle-manager	NS openshift-operator-lifecycle-manager	Succeeded	Apr 29, 2024, 3:17 PM	PackageManifest
Red Hat OpenShift AI 2.8.1 provided by Red Hat	NS redhat-ods-operator	All Namespaces	Succeeded Up to date	Apr 29, 2024, 3:17 PM	Data Science Cluster DSC Initialization FeatureTracker
Red Hat OpenShift Serverless 1.32.1 provided by Red Hat	NS openshift-serverless	All Namespaces	Succeeded Up to date	Apr 29, 2024, 3:18 PM	Knative Serving Knative Eventing Knative Kafka
Red Hat OpenShift Service Mesh 2.5.1-0 provided by Red Hat, Inc.	NS openshift-operators	All Namespaces	Succeeded Up to date	Apr 30, 2024, 11:54 AM	Istio Service Mesh Control Plane Istio Service Mesh Member Istio Service Mesh Member Roll

Red Hat OpenShift

kube:admin

Administrator

Home

Operators

OperatorHub

Installed Operators

Workloads

Serverless

Networking

Storage

Builds

Observe

Compute

User Management

Administration

Project: All Projects

Installed Operators

Name | Search by name...

Name	Namespace	Managed Namespaces	Status	Last updated	Provided APIs
Habana AI 1.15.0-479 provided by Habana Labs Ltd.	NS habana-ai-operator	NS habana-ai-operator	Succeeded Up to date	Apr 30, 2024, 6:10 PM	Device Config
Kernel Module Management 2.1.0 provided by Red Hat	NS openshift-kmm	All Namespaces	Succeeded Up to date	Apr 30, 2024, 11:54 AM	PreflightValidation PreflightValidationOCP Module NodeModulesConfig
LVM Storage 4.14.4 provided by Red Hat	NS openshift-storage	NS openshift-storage	Succeeded Up to date	Apr 29, 2024, 3:17 PM	LVMCluster
Node Feature Discovery Operator 4.14.0-202404161544 provided by Red Hat	NS openshift-nfd	NS openshift-nfd	Succeeded Up to date	Apr 30, 2024, 6:10 PM	NodeFeatureDiscovery NodeFeatureRule
Package Server 0.0.1-snapshot provided by Red Hat	NS openshift-operator-lifecycle-manager	NS openshift-operator-lifecycle-manager	Succee		
Red Hat OpenShift AI 2.8.1 provided by Red Hat	NS redhat-ods-operator	All Namespaces	Succee Up to date		
Red Hat OpenShift Serverless 1.32.1 provided by Red Hat	NS openshift-serverless	All Namespaces	Succee Up to date		
Red Hat OpenShift Service Mesh 2.5.1-0 provided by Red Hat, Inc.	NS openshift-operators	All Namespaces	Succeeded Up to date	Apr 30, 2024, 11:54 AM	Istio Service Mesh Control Plane Istio Service Mesh Member Istio Service Mesh Member Roll

Operators are necessary for Gaudi® to run properly on the Red Hat ® OpenShift platform.

Red Hat OpenShift AI

kube:admin

- Applications
- Data Science Projects
- Data Science Pipelines
- Model Serving
- Resources
- Settings
  - Notebook images
  - Cluster settings
  - Accelerator profiles
  - Serving runtimes
  - User management

Serving runtimes

Manage your model serving runtimes.

Single-model serving enabled | Multi-model serving enabled

Add serving runtime

Name	Enabled	Serving platforms supported	API protocol
Text Generation Inference on Habana Gaudi	✓	Single-model	REST
Caikit TGIS ServingRuntime for KServe Pre-installed	✓	Single-model	REST
OpenVINO Model Server Pre-installed	✓	Single-model	REST
OpenVINO Model Server Pre-installed			
TGIS Standalone ServingRuntime for KServe Pre-installed			

To accelerate your OpenShift AI model with Intel® Gaudi® 2, you need a suitable Serving runtime

Red Hat OpenShift AI

kube:admin

- Applications
- Data Science Projects
- Data Science Pipelines
- Model Serving
- Resources
- Settings
  - Notebook images
  - Cluster settings
  - Accelerator profiles
  - Serving runtimes
  - User management

Accelerator profiles

Manage accelerator profile settings for users in your organization

Name | Find by name | Create accelerator profile

Name	Identifier	Enable	Last modified
Gaudi 2	habana.ai/gaudi	✓	9 days ago

and an adequate Accelerator profile.

Red Hat OpenShift AI

kube:admin

- Applications
- Data Science Projects
- Data Science Pipelines
- Model Serving
- Resources
- Settings
  - Notebook images
  - Cluster settings
  - Accelerator profiles
  - Serving runtimes
  - User management

Data Science Projects > Red Hat Summit LLM w/ RAG Demo

Red Hat Summit LLM w/ RAG Demo

Components | Permissions

Jump to section

- Workbenches
- Cluster storage
- Data connections
- Models and model servers

Workbenches

Create workbench

Name	Notebook image	Container size	Status	
wb1	HabanaAI	Small	Running	Open

Cluster storage

Add cluster storage

Name	Type	Connected workbenches
wb1	Persistent storage	wb1

Data connections

Add data connection

Name	Type	Connected workbenches
bucket-70b-chat	Object storage	No connections
minio	Object storage	No connections

Models and model servers

Deploy model

Single-model serving enabled

Model name	Serving runtime	Inference endpoint	API protocol	Status
tgi-70b-1	Text Generation Inference on Habana Gaudi	https://tgi-70b-1-rag-on-gaudi.apps.cluster1.gaudi.internal	REST	
tgi-70b-2	Text Generation Inference on Habana Gaudi	https://tgi-70b-2-rag-on-gaudi.apps.cluster1.gaudi.internal	REST	

Edit

Delete

Chat Telemetry

What's on your mind?

Stock Llama2 70b

intel xeon intel gaudi

Stock Llama2 70b with RAG

intel xeon intel gaudi redis

What is Red Hat Summit?

Stock Llama2 70b

Hi!

Hello! What can I help you with today?

intel xeon | intel gaudi

Stock Llama2 70b with RAG

Hi!

Hi! Welcome to Red Hat Summit 2024! We're excited to have you join us for this year's event. Can we help you with something specific? Are you looking for information on a particular topic or session? Or perhaps you'd like to know more about our keynote speakers or networking opportunities? Let us know, and we'll do our best to assist you. See you soon!

Sources

Red Hat Summit 2024 s...

intel xeon | intel gaudi | redis

Summary

Red Hat | intel.

Key Takeaways

- RAG enhances AI development
- OPEA simplifies AI deployment
- OpenShift AI integrates into DevOps workflow
- Intel Gaudi 3 accelerates AI training and inference

Confidential AI Helps Protect Data & Models In-Use

Utilizing Confidential Computing for Containers with Intel TDX

Hardware-Based Protection of Data In-Use
With Intel Trusted Domain Extensions (TDX)

+

Openshift sandboxed containers

The Confidential Containers project

Confidential Computing is about protecting data in-use.
You do not have to trust the system admins of the providers any longer.

Q&A

Red Hat | intel.

Red Hat Summit

Connect

Thank you

linkedin.com/company/red-hat

facebook.com/redhatinc

youtube.com/user/RedHatVideos

twitter.com/RedHat

Red Hat | intel.

CODRIN BUCUR

Principal AI Specialist Solution Architect

Red Hat EMEA

Bio: As an Principal AI Specialist Solution Architect, Codrin is supporting Red Hat customers and partners in EMEA with their data science, AI/ML and MLOps needs and best practices. Previously, as Architect and TSM in Red Hat Consulting Alps for 7+ years, Codrin has supported customers with their adoption of Red Hat container platform, integration and middleware technologies.

Contact: **cbucur@redhat.com**

https://www.linkedin.com/in/codrin