SOLUTION BRIEF

Customer Service Call Centers
Virtual Voice Assistants

Intel® Xeon® Scalable Processors Power the Avaamo Conversational AI Voice Assistant to Understand Humans

Intel® Xeon® Gold 6140 processors and Intel® AI technologies power right-sized servers for various deployments

INTEL® AI BUILDERS MEMBER

avaamo

Understanding queries is the first step to an intelligent conversation and satisfying human experience, whether human-to-human or human-to-machine. AI-enabled speech analytics is the core of next-generation intelligent assistants for contact centers. But the technology is a complex architecture of natural language understanding (NLU), graphing for knowledge compilation and ingestion, and deep neural network algorithms, with domain-specific intelligence for the areas of expertise the assistant covers.

A Conversational AI-enabled Assistant

The Avaamo (www.avaamo.com) conversational AI platform is designed to make computers understand humans. Avaamo enables large enterprises to deploy high-impact conversational assistants, offering vertical-specific solutions in various regulated industries, such as healthcare, finance, telecommunications, and others.

The AI-enabled system uses a proprietary NLU engine to process complex queries, with a specific focus on reducing false positives. The NLU engine applies syntactic, semantic, and stochastic processing to distill and discover the purpose behind the user's message. It detects the appropriate tone and sentiment of the query, but also uses additional dimensions to make a more contextual determination on the user's tone, such as:

- User conversation history
- Goals achieved
- User feedback
- Accuracy of prediction

The Avaamo extensions to recurrent artificial neural network algorithms are aimed at maximizing accuracy and recall for varying levels of complexity on the dataset. These extensions include multiple dimension reductions for untagged, unstructured data covering both text and speech. This improves detection of false positives as well as domain-based accuracy resolution.

The Avaamo Knowledge Graph ingests a company's knowledge sources (documents, websites, shared sites, and other sources) and instantly enables the virtual assistant to learn and respond to a customer's natural language queries about that knowledge.

Not only does the Avaamo conversational AI platform understand what the user is asking and which system to get that information from, it can dynamically generate a predictive natural language response to present that information.

Workloads Powered by Intel Xeon Gold 6140 Processors

As an end-to-end solution, the Avaamo conversational AI platform has been optimized for Intel® technologies and is built to address even the most basic of customer challenges like lack of time, expertise, or general natural language understanding (NLU)/classification model development knowledge (also known as the traditional cold-start problem) by:

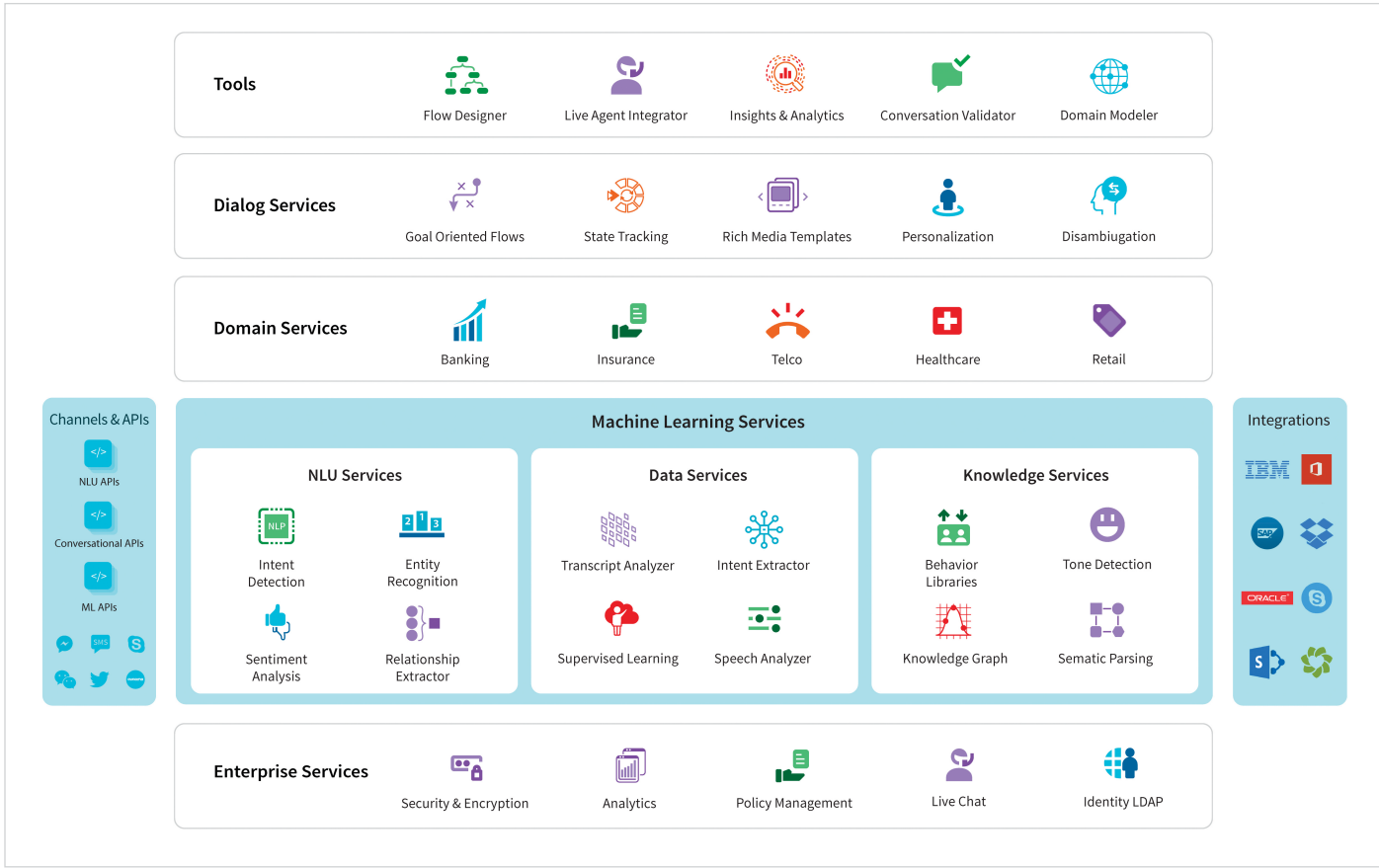
- Ingesting unclassified data
- Performing unsupervised machine learning (ML) model creation
- Optimizing the model for runtime execution
- Enhancing the ML model with customer-specific knowledge resources

Optimized for Intel® Xeon® Scalable processors, each of the following workloads can achieve scalable high- performance with turnaround times dependent on the number of cores available and the size of the workload.

The Avaamo solution, can be deployed on-premise or in the cloud. The following testing was completed on a single server, but the solution easily scales out to accommodate larger workloads. Avaamo can provide details for appropriate configurations for both on-premise and cloud deployments.

Server Configuration

To assess an optimal server configuration for various workloads, a standard production Intel® Xeon® Gold 6140 processor-based dual-socket server was used. Each processor has 18 physical cores running at 2.3 GHz, with two threads per core. This supports a total of 36 virtual cores per socket, an aggregate of 72 virtual cores for the platform. In order to emulate a lower-end, midrange, and high-end server class, the Avaamo workloads were constrained to have core affinity with small (18 virtual cores), medium (36 virtual cores), and high-end server (72 virtual cores) configurations. The tables below illustrate the performance achieved for each workload with different sized deployments.



Results

Ingest and Model Creation Workload

This is the first phase prior to deployment of the inference model. The Avaamo data science automation tool sifts through raw data, like voice/text query transcripts, to understand intent, and then intelligently labels and categorizes the data. From this processed data, using a combination of prebuilt vertical domains and unsupervised training, the Avaamo data science automation tool will create a model for deployment as a virtual assistant that can be further fine-tuned.

This table highlights the approximate duration for a sample workload across the three Intel® Xeon® Scalable processor instances described above.

Transcript Size	**18 cores** (time to completion – rounded to the nearest integer)	**36 cores** (time to completion – rounded to the nearest integer)	**72 cores** (time to completion – rounded to the nearest integer)
64K lines	20 minutes	12 minutes	9 minutes
500K lines	49 minutes	28 minutes	23 minutes
1M lines	124 minutes	72 minutes	60 minutes

Transaction times might vary depending on number of words and complexity within transcript line.

Sizing Guidelines: In most customer contact centers, each agent interaction contains 25 lines of transcript on an average. 1,000 such interactions per day produce 25,000 lines a day. That translates to 9.125 million lines on average in a year.

Server Selection: Depending on your individual performance needs, you can purchase or may already own, a two socket server in a variety of tiers based upon factors including core count per socket, memory configurations, I/O options, etc. Many Original Equipment Manufacturers (OEMs) offer multiple buying options so customers can select the number of cores and sockets that best suits their needs. The table above gives estimates on times to converge to deployment model based on the number of lines of transcript to ingest/process, which can provide guidance to help you select a server tier based on desired completion time that would best fit your needs. This selection process is typically done upfront and offline, before real time deployment, so factors like time to completion are not as critical and the need for special purpose servers is generally not required.

Model Serving and Runtime Execution Workload

The inference model must handle the real-time conversations with a user (employee or customer) with a reasonable amount of latency, so the user experience is not impaired. This test is run with 72 virtual cores. The response times are captured in the table below. The number of requests in aggregate for all sessions is held constant at 10,000 while varying the number of concurrent sessions. Results show that it takes longer to serve more concurrent sessions, even if the total number of requests stays the same.

Concurrent Sessions	**Percentage Complete in Less Than** (seconds – rounded to the nearest integer)		
	90% Complete	**75% Complete**	**50% Complete**
50 sessions 10k requests	4	3	3
100 sessions 10k requests	6	6	5
500 sessions 10k requests	45	31	21

Sizing Guidelines: The 10-10 rule can be applied for most runtime sizing needs. With the 10-10 rule, if you have 10,000 employees, 10 percent of those employees are online, i.e., 1,000. Out of that 1,000, 10 percent will be sending a message at any given time, i.e., 100 concurrent requests. For typical human-to-bot interactions, the users get the best experience with a response time between one and five seconds.

Server Selection: As previously mentioned, the model serving runtime of the virtual assistant can be load balanced across multiple servers with simple front end hardware load-balancer switch. As you scale to more concurrent connections to address a growing client base or more demand, the load-balancer can direct new flow/connection requests to additional servers. The ability to select varying amounts of virtual cores enables you to flexibly grow to any number of servers to address your workload.

Knowledge Graph Building Workload

The Avaamo knowledge graph can ingest company knowledge resources, such as documents and websites, to learn from them and better respond to user queries. Knowledge graphs can vary anywhere from 50 documents to as many as 5,000+ documents in excess of 20 pages each.

Knowledge Data Size (documents)	**36 cores – Time to Completion** (seconds – rounded to the nearest integer)
50	1 minute
500	9 minutes
5,000	115 minutes

Sizing Guidelines: When sizing for Knowledge Graph purposes, it is best to start with an assessment of knowledge dispersion in the enterprise. Some business processes might be cleanly defined with few exceptions and only require 50 or fewer supporting documents to fully enable the end users on how to carry out such business processes. On the other hand, when dealing with processes that extend beyond multiple subsidiaries, or when dealing with customer support processes for 1,000 different product SKUs, the minimum number of documents to start with may be closer to 5,000.

Server Selection: The above table gives initial performance of knowledge graph ingestion from remote networked websites with mid-range server. The performance can be improved with local web server content if done on site with the given end customer data. In addition to variable results due to network congestion, the CPU core utilization was low in this benchmark test, so it is expected that the results can be greatly improved through further core utilization efforts. Given this is not a real-time workload, a mid-range server should be sufficient.

Conclusion

Unlike most traditional training workloads in AI, which require a sizeable investment in specialized hardware with thousands of cores, the Avaamo Model Creation, Model Serving, and Knowledge Graph workloads can function optimally on standard Intel® Xeon® processors and can be scaled gracefully to accommodate larger workloads. This provides immense flexibility for large enterprises to share powerful general purpose Intel hardware across standard and AI-specific computing workloads.

Testing Configurations

In addition to the above processes, Intel® Xeon® Scalable processors were also used for retraining Avaamo's segment specific pre-trained models to meet enterprise customer needs: Testing done December 10-20, 2018

- Linux* OS: Ubunto* 16.04.4 LTS
- Server Hardware: S2600WF, 250 GB Boot drive, 384 GB RAM, Intel® Xeon® Gold 6140 processors (2 socket server, 18 cores each, 36 vCore at 2.3 GHz)
- Intel® Xeon® Processors using latest production S2600WF BIOS/BMC/FRU and are Spectre/Meltdown compliant
- Primary access – routed 10 GbE networking

Avaamo is a member of the Intel® AI Builders Program, an ecosystem of industry-leading independent software vendors (ISVs), system integrators (SIs), original equipment manufacturers (OEMs), and enterprise end users, which have a shared mission to accelerate the adoption of artificial intelligence across Intel® platforms.

0519/BA/HBD/PDF ♻ Please Recycle 338780-001US