SOLUTION BRIEF

Enterprise Communications
AI-Guided Transactions

Real-Time Voice Transcription for Enhanced Customer Experiences

Avaya and Intel team up to transform customer service using AI-powered workflow and experiences

"The processing demands of real-time speech transcription, rules-based workflow, and AI guidance are successfully met by the Intel® Xeon® Scalable processor family and the optimized AI frameworks we rely on to improve the efficiency and scalability of our Avaya Conversational Intelligence solution."
- Jon Alperin, Director AI, Avaya

Customer Strategic Challenges

Artificial intelligence (AI) is rapidly gaining momentum as a strategic tool for enhancing business processes and adding efficiency to organizational workflows. Headquartered in Santa Clara, California, Avaya has established a strong position in the business communication sector by combining AI technologies, cloud computing, and mobile communications to modernize contact center operations around the globe. Recent Gartner Magic Quadrant reports listed Avaya as a leader in unified communications (UC) and contact center infrastructure (CCI).[1]

By integrating AI into its contact center portfolio, Avaya effectively delivers responsive, relevant customer interactions. Intel complements this capability by providing hardware platforms well-suited to auto parallelization and optimized AI frameworks, streamlining and strengthening inference processes and deep learning operations.

The ACI solution enhances customer experiences in the contact center in a number of significant ways:

- Customer calls are transcribed in real time, giving the agent an ongoing view of the items being discussed for easy reference.
- Using AI techniques, the agent receives tips and guidance onscreen to help determine the best way to respond to the customer's needs.
- The customer's mood and sentiments during the call are analyzed by AI—also in real time—to alert the agent (and supervisor) if there are any issues to resolve for the customer.
- After the call concludes, the collected data can be automatically sent to the customer relationship management system (CRM). This reduces agent after-call work by up to 50 percent and provides a record of the customer contact that helps meet regulatory requirements.
- Because all voice conversations are recorded, Avaya Conversational Intelligence can help organizations comply with internal and external rules and regulations.

ACI consolidates customer conversations, assesses trends across thousands of calls, and applies learning algorithms using natural language processing (NLP) to deeply understand customer interactions and refine business operations.

Recently, Avaya re-engineered its ACI solution in collaboration with Intel to enhance the efficiency of automatic speech recognition (ASR) routines. The team effort focused on increasing scalability for handling massive numbers of calls and reducing latency when responding to customer requests in real time. The testing conducted by Avaya demonstrated that by shifting to Intel® Math Kernel Library (Intel® MKL) from OpenBLAS* and migrating to an Intel® Xeon® Platinum 8124M processor-based platform generated a combined 3.75X performance boost when performing ASR operations (as compared to the baseline platform).

ACI Features and Capabilities

The ACI solution is part of the Avaya OneCloud* offerings—a portfolio that includes private, public, and hybrid cloud solution deployments. These solutions cover a wide range of areas, including contact center, company-wide communications processes, and collaboration among geographically separated team members and partners. AI is built into many of these product families.

The deployment of the ACI solution for test purposes relied on Amazon Web Services* (AWS*), although the solution can be deployed on platforms from a number of web service providers. Three distinct AWS instances were employed during the testing, as documented in the configuration end notes. ACI uses open application programming interfaces (APIs) to enable easy interconnection and integration with other applications, including AI-based apps and analytical tools.

Agent recommendations are displayed onscreen during calls using a combination of caller identity data, the real-time automated transcription of the conversation, AI-enabled workflow guidance, and input from the Avaya Business Rules Engine. Kaldi*, an open source application, performs the speech recognition and the TensorFlow* framework was used to develop the deep learning aspects of the solution.

Typical contact center applications that employ speech transcription provide information only after the call has been completed. In comparison, the real-time aspects of ACI not only transcribe the conversations in real time, but actively use AI to flash alerts and suggestions during the call, as shown in Figure 1.

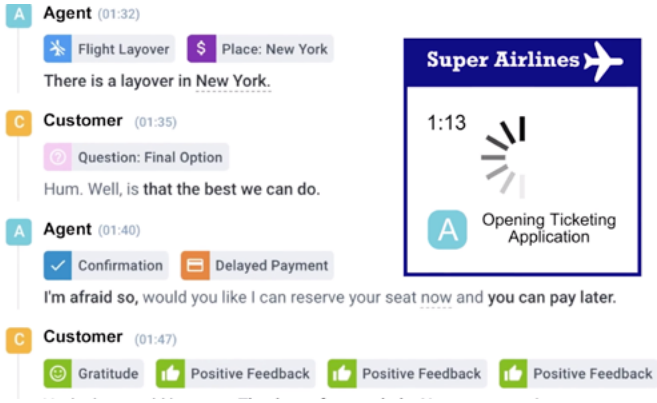

Figure 1. Data and alerts displayed by ACI during a call.

Commentary during the call indicates information related to the call, the customer's frame of mind based on the responses, and can, for example, open a ticketing application at the appropriate point to finalize the transaction. A video on Avaya TV shows the process from the agent's point of view.

AI is an integral component of many different aspects of Avaya contact center solutions and processes, as shown in Figure 2.

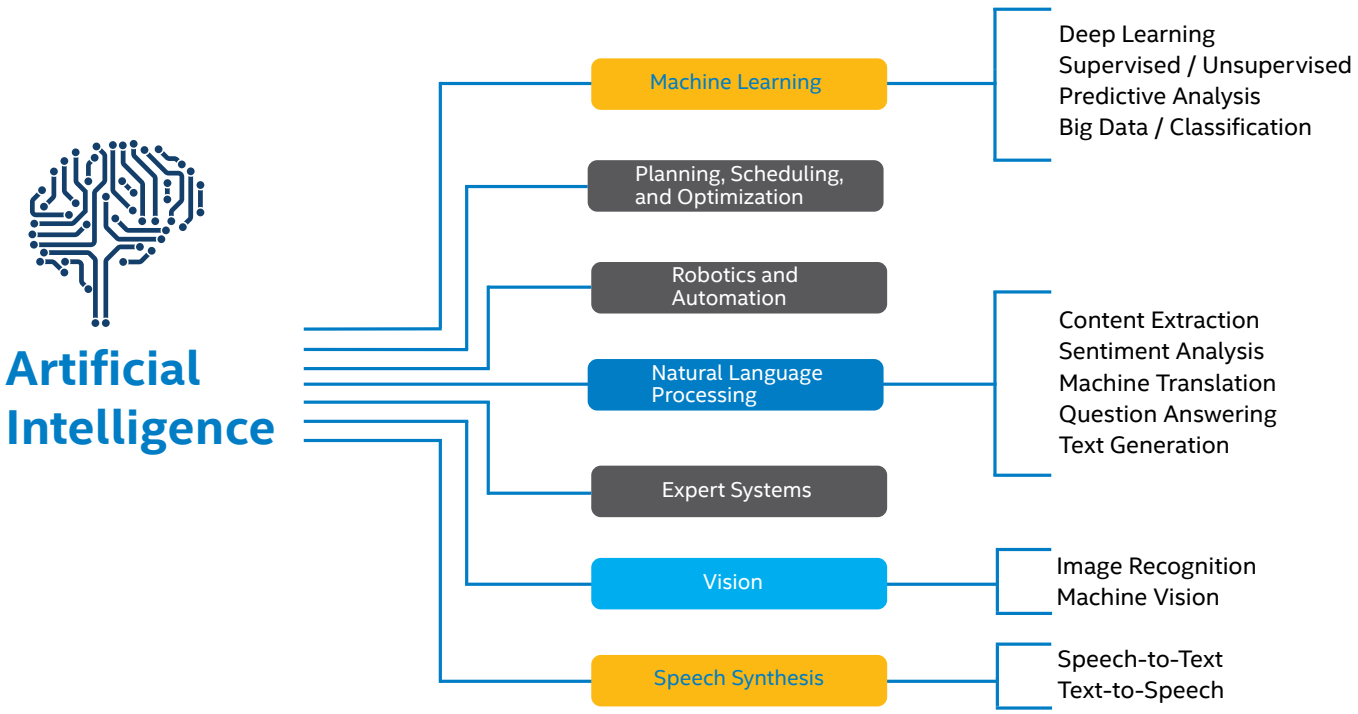

Figure 2. Uses of AI in Avaya contact center solutions.

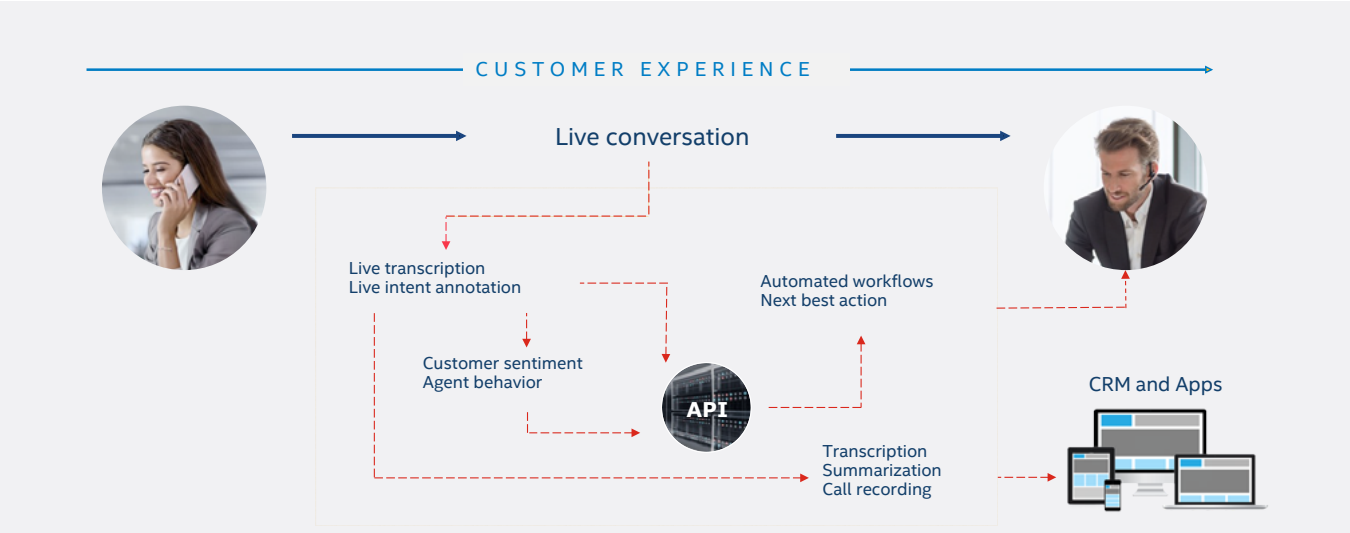

Figure 3. Avaya Conversational Intelligence high-level architecture.

Solution Architecture

Figure 3 shows the high-level architecture of ACI, illustrating the processes involved in using AI to extract intelligence from each call.

Once the live conversation begins, ACI creates a transcription in real time and determines the customer's intent, which is forwarded to an API along with sentiment analysis. Based on business rules and AI input, ACI determines the next best action and forwards this information to the agent.

As shown in the figure, the transcription can be sent to the CRM system or other appropriate application, where it serves as a permanent record of the call linked to customer data.

Building the Intelligent Contact Center

Real-time handling of multiple streams of incoming customer calls—while performing analysis and transcription on the fly—requires substantial processing power, low-latency responsiveness, and enough dynamic scalability to accommodate peak contact center demands that can sometimes reach thousands of calls being processed concurrently.

Moving from OpenBLAS* 0.3.5 to Intel MKL for optimization of algebra functions and migrating to a platform powered by the Intel Xeon Platinum 8124 processor gives Avaya better performance and ensures responsive, real-time call processing on the ACI platform. Intel® Optimization for TensorFlow*, which is integrated into the solution through the optimized libraries provided by Intel MKL 2019, accelerates the deep learning algorithms Avaya developed using the TensorFlow framework.

Intel MKL 2019 features deep-learning primitives optimized for Intel® architecture-based platforms, offering these inherent benefits:

- Refactored code uses modern, efficient vector instructions, ensuring that key operations are vectorized to support the latest Intel® Advanced Vector Extensions 2 (Intel® AVX2) for high performance on Intel® Xeon® Scalable processors.
- Linking AI frameworks with Intel MKL makes it possible to capitalize on the hardware and software optimizations of Intel Xeon Scalable processors without rewriting solution code.
- Built-in data management techniques—including prefetching, cache blocking techniques, and data formats congruent with spatial and temporal locality—ensure that data is available as needed by the execution units.
- Efficient use of available cores can be achieved automatically through Intel MKL, obtaining high levels of parallelization within a given layer or operation as well as across layers.

Intel assistance for enabling the AI capabilities of the solution included both next-generation hardware improvements and the added efficiencies provided by optimized AI frameworks. The resulting enhancements were validated by testing performed by Avaya on AWS instances.

Optimizations and Results

The baseline parameters for evaluating the test results conducted by Avaya and Intel considered these factors:

- **Accuracy**: The accuracy of speech recognition was tracked during testing to ensure that optimization was not affected.
- **Latency**: This parameter was based on the amount of time required to return a response after a customer's spoken query or comment, evaluated from an ASR perspective.
- **Scalability**: Performance was gauged based on the number of ASR processes that could be run on a single processor core, scaled linearly. In other words, if two ASR processes can be run on a single core, testing validated whether eight ASR processes can be run on a four-core processor.
- **Processor utilization**: Test evaluations measured the level of processor performance achieved running one ASR process on a single core.

Test results demonstrated that the use of the Kaldi toolkit with OpenBLAS caused serious performance bottlenecks.[2] Due to the scalable and parallel nature of the workload, switching from OpenBLAS 0.3.5 to the Kaldi toolkit with Intel MKL 2019 immediately boosted the linear scale performance by 2.5 times[3], as shown in Figure 4. Individual ASR processor usage was reduced by 40 percent and latency reduced substantially, without affecting speech recognition precision. This allowed running multiple ASR processes on the same node, increasing system throughput.

Another substantial performance increase was achieved by redeploying the ACI solution—from an AWS instance based on an Intel® Xeon® E5-2666 v3 processor to an instance powered by the Intel Xeon Platinum 8124M processor.[4] This increased scalability by another 1.5 times. These two factors together—the combination of hardware and software improvements—resulted in an overall increase in ASR performance of 3.75x.

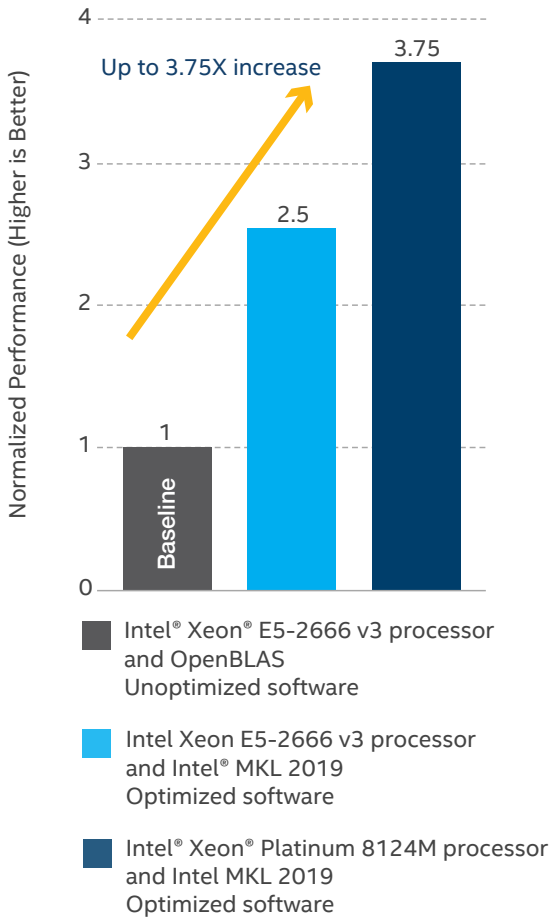

Figure 4. Scalability increases by library and processor.

Ongoing Work and Future AI Strategy

The collaborative work between Avaya and Intel is an excellent example of ongoing explorations into AI and development of practical applications that enable businesses to extract knowledge and useful insights from real-world data. Avaya is a member of the Intel® AI Builders program, an ecosystem that strives to accelerate the adoption of AI by addressing challenges, as well as bringing together the hardware and software ingredients for building solutions. Innovation and compatibility are advanced through this program and the active contributions of members. The successful optimization work on Avaya ASR has opened opportunities for the two companies to explore new ways to use AI and accelerate performance to benefit customers.

Overall, Intel envisions a broad approach to AI[5] that moves beyond the experimental stages to practical applications, shifting from early adopters to early majority. This approach includes:

- Designing specialized chips to support diverse needs in this sector
- Refining existing technologies to advance AI uses
- Acquiring talent to conduct research and develop new hardware and software solutions
- Creating a consistent software layer that intelligently unifies the full range of products

The long-term vision that Intel has for AI includes harnessing deep learning extensively through connectivity to many physical objects across the planet, adding intelligence to a wide range of devices to inspire creativity and innovative solutions for business and personal endeavors. The chipsets and software environments being developed by Intel help solve the challenges of AI and ensure its future effectiveness.

About Avaya

Businesses are built on the experiences they provide, and every day millions of those experiences are built by Avaya (NYSE: AVYA). For over one hundred years, Avaya has enabled organizations around the globe to win—by creating intelligent communications experiences for customers and employees. Avaya builds open, converged, and innovative solutions to enhance and simplify communications and collaboration—in the cloud, on-premises, or a hybrid of both. To grow your business, Avaya is committed to innovation, partnership, and a relentless focus on what's next. As a technology company you can trust, Avaya helps you deliver Experiences that Matter.

Learn More

For more information about the ways that Avaya uses AI in the contact center, visit **www.avaya.com/en/products/contact-center/artificial-intelligence/**

Avaya Conversational Intelligence: www.avaya.com/en/videos/avaya-conversational-intelligence/1_1xh09nax/

Intel AI Builders: builders.intel.com/ai

End Notes

[1] Chirico, Jim. *Avaya Named a Leader in Gartner Unified Communications Magic Quadrant*. Avaya Blog. July 2018. https://www.avaya.com/blogs/archives/2018/07/gartner-uc-magic-quadrant.html.

[2] **BASELINE**: Tested by ISV as of 04/04/2019. AWS* C4.8XLarge Instance: Intel® Xeon® E5-2666 v3 processor @ 2.90 GHz, 2 sockets, 9 cores per socket, 18 cores/36 threads, HT On Turbo ON Total Memory 60 GB, BIOS: 4.2.amazon (ucode: 0x41), Ubuntu* 18.04.2 LTS Deep Learning Framework: Kaldi*, OpenBLAS version: 0.3.5, ARPA Language Model (https://cmusphinx.github.io/wiki/arpaformat/) 40 instances per VM.

[3] Tested by ISV as of 04/04/2019. AWS* C4.8XLarge Instance: Intel® Xeon® E5-2666 v3 processor @ 2.90 GHz, 2 sockets, 9 cores per socket, 18 cores/36 threads, HT On Turbo ON Total Memory 60 GB, BIOS: 4.2.amazon (ucode: 0x41), Ubuntu* 18.04.2 LTS Deep Learning Framework: Kaldi*, Intel MKL version: 2018.1.163, ARPA Language Model (https://cmusphinx.github.io/wiki/arpaformat/) 40 instances per VM.

[4] **NEW**: Tested by ISV as of 04/04/2019. AWS* C5.9XLarge Instance: Intel® Xeon® Platinum 8124M processor @ 3.00 GHz, 1 socket, 18 cores/36 threads, HT On Turbo ON Total Memory 70 GB, BIOS: Amazon EC2 1.0 (ucode: 0x200005a), Ubuntu* 18.04.2 LTS Deep Learning Framework: Kaldi*, Intel MKL version: 2018.1.163, ARPA Language Model.

[5] Green, Tristan. *Intel's AI Strategy for 2019 goes beyond chips*. The Next Web. November 2019. https://thenextweb.com/artificial-intelligence/2018/10/30/intels-ai-strategy-for-2019-goes-beyond-chips/.

[6] Linask, Erik. *Avaya: The Future of Work is All About AI, Mobility, and Other Emerging Tech*. Call Accounting. January 2019. https://call-accounting.tmcnet.com/articles/440957-avaya-future-work-all-ai-mobility-other-emerging.htm.

Notices and Disclaimers

 0919/RA/MESH/PDF 338130-001US