# PART I: PLANNED WORK

(Part I required for a [Planned Work Review](https://w.amazon.com/bin/view/Tech_Promo/Develop_the_Best/Planned_Work_Review/).)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Employee Information** | | | | |
| **Employee Name:** | Kai Zhen | **Current Job Title:** | Applied Scientist II |
| **Manager Name:** | Hieu Duy Nguyen | **Proposed Job Title:** | Applied Scientist III |
| **Steam Member:** | Dave Limp | **Current Business Title:** | Applied Scientist |  | |
| **Steam Direct:** | Tom Taylor | **Proposed Business Title:** | Sr. Applied Scientist |  | |  |
|  | | | | |
| **Goal (e.g., Project, narrative, presentation, DEMONSTRATE A LEADERSHIP PRINCIPLE, etc.)** | | | | |
| *Use your* [*Role Guideline*](https://inside.amazon.com/en/Employment/Career/Role_Guidelines/Lists/Role%20Guideline%20Directory/AllItems.aspx) *to define a single goal that allows you to demonstrate the type of work expected at the next level, our Leadership Principles, or other area to be considered for a promotion. Summarize in one or two paragraphs.* | | | | |
| Reducing runtime latency without compromising accuracy significantly improves customer experience, a pivotal factor for Alexa's market competitiveness. This initiative aims to develop sub-8-bit quantization for compressing ACE chip-enabled on-device ASR (Laser/Theia) with an extra target to extend on-device ASR quantization innovations to cloud ASR scenarios for Alexa technology unification. The primary goal is to decrease Pryon engine latency (P50/90/99) while preserving accuracy, directly linked to Kingpin ([257984](https://kingpin.amazon.com/#/items/257984), [269471](https://kingpin.amazon.com/#/items/269471), [284144](https://kingpin.amazon.com/#/items/284144)). The specific goals are (a) designing and implementing a computation paradigm for sub-8-bit quantization aware training which learns model parameters in a compressed, quantized state; (b) coordinating with ACE team to develop the hardware SDK to leverage the quantized model for acceleration; (c) releasing the first en-US Laser/Theia ASR model in sub-5-bit; (d) mentoring new hires to adopt this tech innovation into more locales in multiple NNA enabled EFD Edge programs; and (e) aligning with Alexa EU team to unify the tech innovation between on-device and cloud scenarios. The goals are measurable in the following way: (a) the model size for Cannoli/CheeseCake needs to be around or under 30MB to unblock the release of the next generation of Echo Dot products; (b) to achieve 15% relative latency reduction measured via Pryon engine latency (P50/90/99); and (c) the accuracy degradation in terms of WER on QBR test sets from quantized models should be less than 1.5% relative. To make the goal actionable, one should (a) be adept in the existing model release procedure in terms of training, packaging, end-to-end accuracy testing, physical-device latency benchmarking, etc; (b) consider building novel algorithms on existing neural efficiency toolsets; and (c) well track the design doc, experimental plan in quip docs and wiki pages. To keep the goal realistic, in terms of implementation and timeline, one needs to (a) work closely with intra-team / cross-team peers and TPM; (b) timely update the progress and risk in the LRs; and (c) deliver promised artifacts in time to build consensus early on. To exhibit the qualities of the next level, one needs to solve problems that are not well defined or structured; build consensus early on among the ASR-PIT team, ASR-EU team, ACE team, and NeMoRT team; work and deliver with limited guidance; and optimize connected systems using their dynamics. | | | | |
|  | | | | |
| **Associated Next-Level Criteria and/or Leadership Principles** | | | | |
| *Cut and paste the next-level criteria or Leadership Principles that this work intends to demonstrate. A single goal can demonstrate readiness in one or more areas. If the work is to demonstrate a different requirement for promotion, document it here.* | | | | |
| **Associated Next Level Criteria:**   * You have significant knowledge/expertise in one or multiple applied science disciplines. * You are able to influence the technical (scientific and engineering) strategy of teams. Understand that not all problems are new (or require new algorithms). * You build and own ML solutions that are easy for others to contribute to. You know how to document solutions and make them auditable, available, and accessible. * You take a long-term view of the business objectives, a system-wide view of the product roadmap, technologies, and how they should evolve.   **Leadership Principles:**   * **Ownership:** Leaders are owners.They think long term and don’t sacrifice long-term value for short-term results. They act on behalf of the entire company, beyond just their own team. They never say “that’s not my job.” * **Invent and Simplify:** Leaders expect and require innovation and invention from their teams and always find ways to simplify. They are externally aware, look for new ideas from everywhere, and are not limited by “not invented here.” As we do new things, we accept that we may be misunderstood for long periods of time. * **Dive Deep**: Leaders operate at all levels, stay connected to the details, audit frequently, and are skeptical when metrics and anecdote differ. No task is beneath them. * **Bias for Action:** Speed matters in business. Many decisions and actions are reversible and do not need extensive study. We value calculated risk taking. | | | | |
|  | | | | |

# PART II: Results

(Complete this section after you finish the work. Parts I, II, and III required for [Completed Work Reviews](https://w.amazon.com/bin/view/Tech_Promo/Develop_the_Best/Completed_Work_Review/).)

|  |
| --- |
| **Quality of Work/Challenge** |
| *Summarize how your work demonstrated the* ***challenge*** *expected at the next level. Where possible, give examples.* |
| **Ambiguity:**  As of 2018, the AHS-ASR team successfully implemented 8-bit quantization aware training (QAT) in AHE-Lite, a pivotal step for the launch of Raven, the Fire TV Cube 2nd generation device. However, transitioning from 8-bit to sub-8-bit (5-bit) quantization posed unexpected challenges. While employing 5-bit quantization reduces NNA bandwidth by 37.5%—significantly enhancing user-perceived latency crucial for Alexa’s customer experience—it heavily restricts model weight dynamics and capacity, raising accuracy concerns. Unlike 8-bit quantization with fixed linear centroids within a closed interval, determining optimal locations for 32 centroids in 5-bit quantization is essential for maintaining viable accuracy. Simply applying the previous 8-bit QAT method to 5-bit models resulted in an accuracy decrease of over 10%. The debate persists on how to effectively compress on-device ASR models to 5-bit using a reusable end-to-end solution. In summary, sub-8-bit quantization remains widely perceived as an inherently complex and ambiguous optimization challenge.  Secondly, in the realm of ASR model quantization, the complexity rises, particularly in cross-team Software/Hardware Co-Design. Whether 8-bit or 5-bit quantization, it makes no difference to performance if the underlying hardware can’t take advantage of them For on-device ASR utilizing NNA, the critical question arises: will the 5-bit model execute faster? In this initiative, I initiated feature request intake tickets in the project's early stages, directing them to the ACE team. This move aimed to promptly verify hypotheses and invest in promising co-design avenues. Tacking the ambiguity with this mechanism, we observed the frame processing rates for the 5-bit baseline to be 17.1% faster than the 8-bit-compressed models. Simultaneously, we advocated for implementing the Lloyd-Max algorithm in 5-bit quantization to optimally position 32 quantized weight centroids. Consequently, the ACE team, aligning with our in-training Lloyd-Max algorithm, enabled post-training quantization via the K-means clustering algorithm at runtime to process our 5-bit ASR models. This successful Software/Hardware Co-Design venture yielded zero accuracy degradation when compressing models to 5-bit.  Last but not least, in extending our on-device ASR quantization to cloud scenarios, we encountered limited guidance, primarily due to the inherent project ambiguity, leaving critical questions unanswered. For example, has Cloud Conformer already been executed in 8-bit on the CPU? If so, how are weights quantized in Conformer? Will the existing post-training 8-bit weight conversion approach be as good as our training-aware General Quantizer (GQ) in preserving the accuracy? What are the boxes to be checked to test our ASR quantization solution end-to-end in the current Cloud-Conformer-based production pipeline, and so on? Despite these challenges, persistent collaboration with partner teams and cloud ASR release owners allowed us to address these uncertainties. Collaborating closely with NeMoRT (Chris Beauchene, Sr. SDE), we identified discrepancies between model training in 32-bit and execution on CPU in 8-bit, clarifying how this discrepancy affected quantization-induced accuracy loss in cloud ASR models, particularly in WBR test sets. Working with the ASR-EU team (Jahn Heymann, Sr. Applied Scientist), we rigorously tested GQ at various stages, including core-transducer training, checkpoint averaging, incremental learning, and neural biasing. Through concerted efforts, we were able to deploy GQ to cloud de-DE ASR with 1-3% WERR on Glidepath and Tail, mitigating previous quantization-induced losses. Subsequently, GQ was implemented in en-GB and en-AU. Despite the lack of guidance, I took ownership, surmounted obstacles and production challenges, navigated through the ambiguous process, and ultimately completed the project in the Runtime Modeling program.  **Scope of Influence:**  Influencing other teams presents challenges, particularly when introducing on-device innovations to cloud teams, as usually, the influence flow is the other way around. To overcome this, I proactively scheduled 1:1 meetings and developed questionnaires to gather inputs and address concerns from cloud ASR and NeMoRT teams (Hitesh Tulsiani, Jahn Heymann, Harish Arsikere, Chris Beauchene). Iterating on their feedback, I proposed potential solutions backed by detailed data and figures for further discussions. Engaging software/hardware experts from multiple teams, we gradually established consensus that integrating 8-bit GQ into cloud Conformer core-transducer training is an achievable objective within the Runtime program. This integration holds significant value in enhancing model accuracy especially on Glidepath and Tail test sets.  To ensure timely releases for on-device ASR (Bluebottle/Crosstown), I collaborated closely with the ACE team (Raviteja Chinta, Sr. SDE) to co-design the 5-bit NNA SDK. Additionally, I provided active mentorship to scientists involved in productizing 5-bit quantization. For instance, I facilitated the onboarding of Yi Xie (Applied Scientist II, AHS-ASR team) to lead the productization of 5-bit quantization in the en-US Crosstown model. Moreover, I developed Runbooks and crafted a video tutorial on ODIE/Crosstown TVM packaging to mentor Rohit Barnwal (Applied Scientist I, AHS-ASR team, now at Tiktok) for the release of the en-GB Crosstown model. As a result, our team successfully integrated 5-bit quantization across all locales for Brownie/Ganache and Cannoli/CheeseCake. With 5-bit, the size of our Crosstown models is reduced by over 30%, which successfully resolved the business pain point caused by the memory bottleneck on NNA-v2 for Cannoli/CheeseCake. For Stage1 Pryon Latency (ms) on NNA v1, 5-bit quantization achieved 373.00 msec from 788.00 msec, or over 50% user-perceived latency reduction, which is a huge win for customers’ experience.  In summary, shaping aligned consensus across multiple teams isn't a consequence but a prerequisite for driving 5-bit quantization in on-device ASR (Bluebottle/Crosstown) and introducing 8-bit General Quantization to cloud ASR. Additionally, I actively showcase our research and development progress on 5-bit quantization across various internal forums, including the wake-word team meeting, core-transducer workstream LR, and the AMLC workshop.  **Scientific and Technical Complexity:**  Sub-8-bit quantization was rarely investigated as they require sub-8-bit operators on neural network accelerators (NNAs), which often have inferior performance compared to their 8-bit counterparts due to the reduced numerical accuracy. Consequently, sub-8-bit NNAs are less adopted, and thus, there is no real latency measurement for existing sub-8-bit approaches. To leverage sub-8-bit arithmetic on INT8-based NNA, I proposed a novel sub-8-bit quantization aware training (S8BQAT) that was published in Interspeech’22 among the first few papers in this area.  S8BQAT distills a subset of 32 quantization centroids from a pre-trained 32-bit baseline via a mechanism derived from Lloyd-Max scalar quantization theory. I introduced Multi-Regional Absolute Cosine (MRACos) regularizer, which is INT8 compatible and computationally efficient. The MRACos regularizer imposes a probability distribution on the network weights during training by penalizing off-the-centroid weights and aggregating them towards their nearest quantization centroids. Additionally, the MRACos regularizer is accompanied by a periodic compressor that assigns each model weight to that nearest quantization centroid, ensuring quantization convergence and, therefore, minimizing runtime quantization-induced performance degradation. As mentioned before, S8BQAT achieves a superior WER-UPL tradeoff compared to our previous 8-bit QAT baseline. In particular, with S8BQAT, we increase the number of Bluebottle’s model parameters by 10.3%, thus reducing WER by 4-16% relative while reducing UPL by 5%.  While S8BQAT was successfully productized in multiple RNN-T-based on-device ASR programs, we did not stop but insisted on the highest standard. With guidance from Hieu Nguyen (AS manager in AHS-ASR) and Grant Strimel (Pr. Applied Scientist in AHS-ASR), I further invented the next generation of S8BQAT, termed General Quantizer (GQ).  Inspired by S8BQAT, GQ features a regularization-free, “soft-to-hard” compression mechanism with self-adjustable centroids in a µ-Law constrained space, resulting in a simpler yet more versatile quantization scheme. Same with S8BQAT, we observe a 30.73% memory footprint saving and 31.75% user-perceived latency reduction compared to 8-bit QAT via physical device benchmarking. What’s superior is that GQ can compress both RNN-T and Conformer into an arbitrary bit-depth setting with minimal manual effort as a play-and-play solution ASR quantization.  **Impact:** Sub-8-bit/general quantization has made a broad impact in terms of both research and production.  ***For on-device ASR production:*** it effectively lowers the user perceived latency by over 30% ([Kingpin Goal](https://kingpin.amazon.com/#/items/257984) [257984](https://kingpin.amazon.com/#/items/257984)), which affords a larger RNN-T architecture for improving recognition accuracy. I deployed sub-8-bit quantization in Bluebottle (Laser/Theia) en-US R15 ([Wiki page](https://wiki.labcollab.net/confluence/display/SHELBY/BlueBottle+R15+en-US+RNN-T+Release#Project-1430512554)), as the first 5-bit ASR model in Alexa: the model was able to feature 11.7% more parameters to achieve WERR of 15.3% on glidepath and 18.8% on messaging, yet still with reduced latency. Similarly, for CrossTown (with devices of more constrained memory size), I am the release owner of the es-ES model ([Wiki page](https://wiki.labcollab.net/confluence/display/SHELBY/Training+and+Delivery+of+Crosstown+es-ES+v3+ASR+model#0--1532767274)), the first non en-US locale with 5-bit quantization: it reduces the memory footprint by 46.0% from 54MB to 29MB on NNA v2, which solved the memory bottleneck of deploying ASR models to 5th generation of Echo Dot (Cannoli/CheeseCake). I also onboarded a few other applied scientists to productize 5-bit quantization in multiple locales for the BlueBottle/CrossTown program. By the end of 2022, we have deployed sub-8-bit quantization-enabled ASR architectures to all of our NNA-enabled EFD Edge programs (Bluebottle and Crosstown) across all locales, which includes the products Echo, Echo Dot, and Echo Show for Amazon.com Inc. voice-controlled assistants.  ***(make cloud case stronger, consider it first, perception is that bb is not that important)***  ***For cloud ASR production:*** after the deployment of sub-8-bit QAT to on-device ASR, I kept pushing the envelope and brought more innovation to the ASR model in-training quantization, called General Quantizer (GQ), making the recipe more concise and reusable among arbitrary model architectures either on-device or in-the-cloud. GQ further improves the simplicity of enabling quantization-aware training: the length of code is reduced by over 28% in Phasa mainline (GQ task); GQ is model-agnostic, making the QAT mechanism a callback-based, plug-and-play solution. This effectively lowers the endeavor of the release owner to intake our innovation. I deployed GQ in de-DE v59 - the first unified cloud Conformer model [[Wiki page](https://wiki.labcollab.net/confluence/pages/viewpage.action?pageId=2054818930)] that was trained with General Quantization (GQ), which yields 2% WERR on tail (from 7.41% to 7.24%) and 3% WERR on wbr (from 7.62% to 7.41%) (Kingpin Goal) (highlighted in QBR of 2023/Q2). As another example of software-hardware codesign, GQ directly leverages INT8 Hybrid GEMM developed by Alexa Speech Engine team for runtime speedup. Compared to the post-training quantization approach, GQ reduces the mismatch between the training with FP32 precision and runtime with INT8 precision via quantization-aware training (QAT). By the end of 2023, GQ has been productized for cloud Conformer in de-DE, en-GB and en-AU locales.  ***Research impact:*** In collaboration with hardware experts from the ACE team, under the guidance of senior and principal scientists, I have secured the invention via a provisional patent application, such that Alexa ASR wouldn’t risk defending a patent infringement suit had its competitors patented it in the first place.   * ***Kai Zhen****, Hieu Nguyen, Raviteja Chinta, Tariq Afzal, Anastasios Alexandridis, Athanasios Mouchtaris, Ariya Rastrow, Compression of Machine Learned Models, P77898-US01*   I have also published academic papers for sub-8-bit quantization in top-tier conference proceedings, demonstrating Alexa’s leading role in the ASR/ML field with citations from external prestigious research institutes, such as University of Cambridge, Tsinghua University, RWTH Aachen University, Samsung Research UK, and Google Research, etc.   * ***Kai Zhen****, Hieu Duy Nguyen, Raviteja Chinta, Nathan Susanj, Athanasios Mouchtaris, Tariq Afzal, and Ariya Rastrow, "*[*Sub-8-Bit Quantization Aware Training for 8-Bit Neural Network Accelerator with On-Device Speech Recognition*](https://assets.amazon.science/fe/84/ad0cdd7c4967b17aaf670fe0194b/sub-8-bit-quantization-aware-training-for-8-bit-neural-network-accelerator-with-on-device-speech-recognition.pdf)*," In Proc. Annual Conference of the International Speech Communication Association (Interspeech), Incheon, Korea, September 18-22, 2022.* * ***Kai Zhen****, Martin Radfar, Hieu Nguyen, Grant Strimel, Nathan Susanj, Athanasios Mouchtaris, "General Sub-8-Bit Quantization for On-Device Speech Recognition: A Regularization-Free Approach," in Proceedings of the 2022 IEEE Spoken Language Technology Workshop, Doha, Qatar, January 9-12, 2023.*   **Execution:**  I prioritize both intra-team and cross-team coordination, especially during the project design and scoping phase. This approach fosters early consensus-building, resulting in highly adaptable end-to-end solutions. For instance, I proactively engaged with Jahn Heymann (Sr. Applied Scientist in ASR-EU) and Chris Beauchene (Sr. SDE in NeMoRT) to gain insights into the comprehensive release procedure for Conformer-based cloud ASR and its execution on CPU using INT8Hybrid GEMM.  By harmonizing diverse perspectives from scientists and engineers, I devised solutions that seamlessly integrated with the AHE Engine team's ongoing work. Consequently, our published scientific study on "General Quantization (GQ)" was effectively implemented in the de-DE cloud Conformer, leveraging NeMoRT’s INT8 kernel for runtime acceleration. This implementation notably resulted in a 1-3% improvement in WBR/Tail test sets due to reduced post-training quantization-induced loss. Subsequently, Venkata Kishore Nandury (Sr. Applied Scientist in ASR-BLR) successfully deployed GQ to en-GB and en-AU locales for cloud ASR, requiring minimal assistance from me. This successful deployment underscores GQ's adaptability as a comprehensive 8-bit quantization solution for end-to-end implementation.  Compressing on-device RNN-T-based ASR models emerged as a top priority program for AHS-ASR (Bluebottle and Crosstown). Achieving this goal demanded extensive cross-team collaboration, primarily among AHS-ASR, ACE, and Automation, orchestrated by Hieu Duy Nguyen (AS manager in AHS-ASR) and Brian Collins (Sr. TPM in AHS-ASR). I also collaborated closely with Raviteja Chinta (Sr. SDE at ACE, now at Meta) to test 5-bit prototypes on NNA v1 and v2, validating their latency and memory advantages. Additionally, I spearheaded the automation of the 5-bit weight conversion process with the automation team.  Taking the initiative, I devised the runback for converting the NNA v1 model to the NNA v2 compatible version. Through these concerted efforts, we successfully deployed 5-bit quantization-enabled ASR architectures across all locales in 2022, within our NNA-enabled EFD Edge programs (Bluebottle and Crosstown). This implementation led to a reduction of over 30% in the size of our Echo/Echo-dot models without sacrificing accuracy. Consequently, Alexa can now cater to more customers using diverse device types, including those with constrained memory sizes, such as Cannoli/CheeseCake—the 5th generation of Echo Dot.  **Knowledge:**  As previously outlined, my expertise spans a wide spectrum within ASR model compression, evident in achieving objectives for RNN-T-based Bluebottle/Crosstown programs and Conformer-based cloud ASR runtime programs. Notably, our 'sub-8-bit' quantization research stands out as a top-ranked search on Google, drawing citations from esteemed research and industrial institutions.  Successfully productizing these technological innovations demands more than just specialized expertise; it necessitates a comprehensive understanding of critical ASR components like acoustic modeling, language modeling, and end-pointing to devise viable solutions. My proficiency in both model compression and Alexa ASR's technology stack enables swift prototyping and experimental validation—from unit-level in-training (Phasa/PyRama) to service-level end-to-end testing on accuracy (Djinn and DoryBlueshift-ASR) and latency (Physical-Device-Benchmarking-Portal).  This capability has facilitated multiple production deployments for on-device and cloud ASR scenarios, delivering substantial benefits to Alexa's global customer base. Particularly in the era of Generative AI with Large-ASR and NextGen-ASR, these deployments have proven highly advantageous.  **General Qualifications:**  I was first hired by AHS-ASR in May 2020 as an applied scientist intern honored to have my work recognized among the top 17 poster presentations. Later, I joined the AHS ASR group at Amazon PIT full-time, I've spearheaded several impactful neural efficiency projects for both on-device and cloud ASR, leaving a significant mark on production and research initiatives. Prior to Amazon, I contributed to BERT-based multi-modal recommendation systems as a machine learning relevance intern at LinkedIn from 2018 to 2019. I received a B.S. degree in Software Engineering from Xidian University in 2012 and an M.S. in Computer Science from Tsinghua University in 2015, where I won the Honorable Mention award for the International Mathematical Contest in Modeling (MCM) and was a 3-time recipient of the National Scholarship. I obtained a Ph.D. degree in computer science and cognitive science from Indiana University Bloomington, where I led multiple speech and audio neural waveform coding projects, which pioneered a new research area. My papers were published in leading signal processing and speech processing conferences and journals, such as Interspeech, ICASSP, IEEE Signal Processing Letters (SPL), and IEEE T-ASLP. The Cognitive Science Program at IU recognized my research by awarding the Outstanding Research Award in 2021. Furthermore, I hold six US patents as an inventor, reflecting my contributions to innovative technological solutions. |

|  |
| --- |
| **Impact/Measure of Success** |
| *Summarize in one or two paragraphs the impact this effort had on customers, your team/organization/business area, partners, and others. What data (quantitative1 metrics or qualitative measures) can you provide?* |
| **Reducing the accuracy gap between Bluebottle and Doppler ASR models**  S8BQAT, as a Software Hardware Co-Design solution for on-device speech processing, makes Alexa faster, has lower bandwidth, and higher predictive performance. For the Bluebottle program, corresponding to Echo, Echo Show, and FireTV products, reducing the accuracy gap against the cloud model is of high priority (Kingpin goal). As a concrete example, S8BQAT affords the AM of Bluebottle en-US R15 to increase its number of model parameters by 11.7%, which effectively boost the model accuracy on glidepath (from 6.81% to 6.4%, or 6.41% WERR), rare\_words (15.34% to 14.32%, or 7.12% WERR) and contact-name (8.81% to 7.48%, or 17.78% WERR). Although going larger, the model achieves 8.56% p50 UPL reduction (from 1391ms to 1272ms). Persistently pushing from there, in Bluebottle en-US R18, our on-device model has achieved the goal of lowering the accuracy gap against the Cloud counterpart to 15% relatively or less.  **Elevating customers’ experience with significantly lowered model size and user-perceived latency**  For the Crosstown program to handle on-device intent execution, the corresponding device type is Echo Dot, with far more strict constraints on memory footprint. For example, Cannoli/CheeseCake, the 5th generation of Echo Dot, can only shoehorn our on-device ASR models onto the device if its size is around 30MB or less. With S8BQAT-empowered 5-bit compression, we’ve successfully met the goal: for Crosstown en-US R15, the model size in 5-bit goes down from 31 MB to 23 MB, or a 25.81% relative reduction. Similarly, for es-ES, the model size in NNA v1’s anfbin format is lowered by 17.20%, which is accompanied by a 12.52% p50 UPL reduction and 6.31% WERR on glidepath accuracy.  **Unifying quantization tooling between edge and cloud ASR computing**  In the Runtime Modeling program, the next generation of S8BQAT, General Quantizer (GQ), was integrated into cloud ASR as the unified Conformer-based core-transducer release recipe. As highlighted in QBR of 2023/Q2, de-DE v59 became the first unified cloud Conformer model [[Wiki page](https://wiki.labcollab.net/confluence/pages/viewpage.action?pageId=2054818930)] that was trained with General Quantization (GQ), which yields 2% WERR on Tail (from 7.41% to 7.24%) and 3% WERR on WBR (from 7.62% to 7.41%) (Kingpin Goal). Later, in 2023, GQ was productized in en-GB and en-AU locales as well, making our team’s innovation go beyond on-device ASR but cloud ASR as a unified and end-to-end validated neural efficiency tool.  **Securing intellectual property via patent and publications**  From on-the-edge to over-the-air scenarios, the intellectual property on ASR model quantization from both the software and hardware’s perspectives are completely owned by Amazon. With the scientifically sounding solution, successful product integration, and foreseeable business value, we filed the “Method and Apparatus of Sub-8-Bit Quantization-Aware Training for Deep Learning Applications”, which has been approved by Amazon’s legal team. Both S8BQAT and GQ papers are accepted at flagship speech processing international conferences (Interspeech’22 and IEEE SLT’23, respectively).   * ***Kai Zhen****, Hieu Nguyen, Raviteja Chinta, Tariq Afzal, Anastasios Alexandridis, Athanasios Mouchtaris, Ariya Rastrow, Compression of Machine Learned Models, P77898-US01* * ***Kai Zhen****, Hieu Duy Nguyen, Raviteja Chinta, Nathan Susanj, Athanasios Mouchtaris, Tariq Afzal, and Ariya Rastrow, "*[*Sub-8-Bit Quantization Aware Training for 8-Bit Neural Network Accelerator with On-Device Speech Recognition*](https://assets.amazon.science/fe/84/ad0cdd7c4967b17aaf670fe0194b/sub-8-bit-quantization-aware-training-for-8-bit-neural-network-accelerator-with-on-device-speech-recognition.pdf)*," In Proc. Annual Conference of the International Speech Communication Association (Interspeech), Incheon, Korea, September 18-22, 2022.* * ***Kai Zhen****, Martin Radfar, Hieu Nguyen, Grant Strimel, Nathan Susanj, Athanasios Mouchtaris, "General Sub-8-Bit Quantization for On-Device Speech Recognition: A Regularization-Free Approach," in Proceedings of the 2022 IEEE Spoken Language Technology Workshop, Doha, Qatar, January 9-12, 2023.* |
|  |

|  |
| --- |
| **Leadership Skill & InFluence (Section required for promotions into L6+ only)** |
| Please describe how this work gave you the opportunity to demonstrate any other leadership skills (e.g., drive a best practice, influence a needed change, build consensus on an approach)? How did you apply judgment (e.g., design choices, trade-offs, priorities) or make decisions with long-term effects? |
| In this project, I consistently perform at the next level. I addressed the need for a fast and accurate on-device speech recognition model interface by studying model quantization techniques. Via my effort, we found that using a sub-8-bit quantization algorithm allowed for the compression of the model weight from 32-bit to 5-bit without decreasing predictive performance. This resulted in the presentation of a novel sub-8-bit quantization-aware training scheme for 8-bit neural network accelerators (NNA) inspired by the Llyod-Max compression theory with a practical adaptation for a feasible computational overhead during training.  I influenced multiple teams—ACE, ASR-EU, NeMoRT—collaboratively introducing (sub)-8-bit training for both on-device and cloud ASR. Beyond Amazon, I led scientific impact by publishing 2 peer-reviewed papers and filing 1 patent application based on our quantization work, collaborating with co-authors from various organizations.  The science and production impact showcased through this project serves as a testament to my readiness to become a senior applied scientist, bringing more novel methods and innovations into a range of customer products with substantial commercial value and practical applicability in real-world scenarios in the era of efficient generative AI.  **Invent and Simplify:** I maintained a forward-looking perspective, consistently pursuing avenues to enhance and streamline the neural efficiency toolset while overseeing Alexa ASR's shift from RNN-T to a Conformer-based core-transducer architecture. This commitment led to the PIT ASR team's neural efficiency solution becoming seamlessly integrated and well-aligned with the advancements in ASR model architectures over time. These solutions remain indispensable component across multiple ASR programs, including AutoS2I, Large-ASR, and Runtime Modeling, continuing to play a vital role in our ongoing initiatives for Next-Gen Generative AI empowered ASR.  **Bias for Action:** I operated with a bias for action, maintaining continuous communication with various teams to gather diverse insights. Constantly exploring innovative ideas aimed at addressing technological integration challenges for the benefit of our customers, I successfully executed the General Quantizer (GQ) project in Q2/2023, introducing 8-bit quantization-aware training to cloud ASR. It's worth noting that amidst the organization's shift toward Large ASR modeling, the completion of the GQ project stood out among the few accomplishments in Q2. This showcases my ability to take calculated risks and navigate team conflicts, ensuring timely delivery—an invaluable skill in the intricate and often ambiguous cycles of product development.  **Ownership:** As an applied scientist, I embrace diverse responsibilities beyond scientific endeavors. These include facilitating cross-team collaborations, resolving conflicts among research scientists and machine learning engineers, and engaging in production triage discussions with TPMs. Understanding the significance of overseeing every facet of the release procedure, I prioritize the transition of technological innovations into production and the hands of customers. Even after our technology's deployment, I remain proactive in safeguarding the company's intellectual property by initiating the patent filing process. This strategic step aims to fortify Amazon's IP portfolio, ensuring protection against potential competition in the market.  **Dive Deep**: I actively maintain connections across multiple layers within the Alexa ASR's technology stack and engage with cross-organizational teams. This proactive engagement is aimed at cultivating a comprehensive understanding of Alexa's ecosystem. I approach benchmarks with curiosity and skepticism, evident in my investigation of GQ productization in cloud ASR. Despite the prevalence of post-training dynamic quantization, I delved into details, collaborating with NeMoRT's Senior SDEs and ASR-EU's Senior Applied Scientists. This scrutiny unveiled scenarios where quantization-induced loss impacted accuracy, particularly in the presence of weight outliers. This in-depth investigation led me to validate the business viability of implementing GQ for cloud ASR, ultimately resulting in enhanced accuracy for Tail and WBR metrics using NeMoRT's established INT8 HybridGEMM toolset. This underscores the tangible benefits arising from technical exploration and collaboration. |

|  |  |  |
| --- | --- | --- |
| **ARTIFACTS** | | |
| *Provide links below.* | | |
|  | **Artifact Name and Link** | **Description** |
|  | [First 5-bit en-US Bluebottle model release](https://wiki.labcollab.net/confluence/display/SHELBY/BlueBottle+R15+en-US+RNN-T+Release#Project-388125852) | Wiki for sub-8-bit trained en-US Bluebottle R15 release as the first 5-bit on-device ASR model |
|  | [First 5-bit non en-US Crosstown model release](https://wiki.labcollab.net/confluence/display/SHELBY/Training+and+Delivery+of+Crosstown+es-ES+v3+ASR+model) | Wiki for sub-8-bit trained es-ES Crosstown R3 release as the first 5-bit trained non English ASR model |
|  | [First 8-bit cloud Conformer model release (de-DE)](https://wiki.labcollab.net/confluence/pages/viewpage.action?pageId=2054818930) | Wiki for general quantization (GQ) and its intake for cloud ASR (de-DE v59) |
|  | [Patent application inventory on S8BQAT](https://quip-amazon.com/PKW5ALZ7BVhU/Patent-Method-and-Apparatus-of-Sub-8-Bit-Quantization-Aware-Training-for-On-Device-Deep-Learning-Applications) | The application inventory that leads to a filed patent with the title of “*Compression of Machine Learned Models*”, *P77898-US01.* |
|  | [S8BQAT for Interspeech’22](https://www.isca-speech.org/archive/pdfs/interspeech_2022/zhen22_interspeech.pdf) | Our paper with the innovation productized for on-device RNN-T |
|  | [GQ for IEEE SLT’23](https://assets.amazon.science/0c/03/41fc077547799c2350ccb3a4ac15/sub-8-bit-quantization-for-on-device-speech-recognition-a-regularization-free-approach.pdf) | Our paper with the innovation productized for cloud Conformer |
|  | [Launch announcement](https://quip-amazon.com/bTBQAjhiM2KM/Launch-Announcement-8-bit-Cloud-Conformer-Training-via-General-Quantization) | Launch Announcement for 8-bit Cloud Conformer Training via General Quantization |
|  | [Runtime Modeling LR Highlights](https://quip-amazon.com/nykpAZKcR9kV/Runtime-Modeling-Bi-Weekly-Update-07112023#temp:C:TAZ529cda86fbcf4f208965c3b3f) | Highlights on 8-bit Cloud Conformer Training via General Quantization after its deployment |
|  | [Live Latency Monitor](https://monitorportal.amazon.com/igraph?SchemaName1=Service&DataSet1=Prod&Marketplace1=USAmazon%3Abrownie&HostGroup1=ALL&Host1=ALL&ServiceName1=AlexaHybridEngine&MethodName1=FirstPassRecognition&Client1=ALL&MetricClass1=NONE&Instance1=NONE&Metric1=stage1_pryon_latency_msec.es-ES&Period1=OneDay&Stat1=p50&ValueUnit1=microsecond&LiveData1=true&Label1=p50%20-%20OneHour%20%5Bmin%3A%20%7Bmin%7D%2C%20avg%3A%20%7Bavg%7D%2C%20max%3A%20%7Bmax%7D%5D&Color1=1600ff&Visible1=false&UserLabel1=p50%20-%20OneHour%20%5Bmin%3A%20%7Bmin%7D%2C%20avg%3A%20%7Bavg%7D%2C%20max%3A%20%7Bmax%7D%5D&SchemaName2=Service&Stat2=p90&Label2=p90%20-%20OneHour%20%5Bmin%3A%20%7Bmin%7D%2C%20avg%3A%20%7Bavg%7D%2C%20max%3A%20%7Bmax%7D%5D&Color2=0a0a0a&UserLabel2=p90%20-%20OneHour%20%5Bmin%3A%20%7Bmin%7D%2C%20avg%3A%20%7Bavg%7D%2C%20max%3A%20%7Bmax%7D%5D&SchemaName3=Service&Stat3=p99&Label3=p99%20-%20OneHour%20%5Bmin%3A%20%7Bmin%7D%2C%20avg%3A%20%7Bavg%7D%2C%20max%3A%20%7Bmax%7D%5D&Color3=ff00ff&Visible3=true&UserLabel3=p99%20-%20OneHour%20%5Bmin%3A%20%7Bmin%7D%2C%20avg%3A%20%7Bavg%7D%2C%20max%3A%20%7Bmax%7D%5D&SchemaName4=Service&Period4=OneHour&Stat4=n&ValueUnit4=millisecond&LiveData4=false&YAxisPreference4=right&Label4=n%20-%20OneHour%20%5Bmin%3A%20%7Bmin%7D%2C%20avg%3A%20%7Bavg%7D%2C%20max%3A%20%7Bmax%7D%5D&Color4=02e0e0&Visible4=false&UserLabel4=n%20-%20OneHour%20%5Bmin%3A%20%7Bmin%7D%2C%20avg%3A%20%7Bavg%7D%2C%20max%3A%20%7Bmax%7D%5D&HeightInPixels=720&WidthInPixels=1560&GraphTitle=stage1_pryon_latency_msec&DecoratePoints=true&GraphType=zoomer&HorizontalLineLeft1=%23color%3Dblue%20Sev2%20%28p50%29%20-%20@%201920.58&HorizontalLineLeft2=%23color%3Dgreen%20Sev2%20%28p90%29%20-%20@%204206.02&HorizontalLineLeft3=%23color%3Dmagenta%20Sev2%20%28p99%29%20-%20@%209432.25&HorizontalLineLeft4=%23color%3Dblue%20Sev3%20%28p50%29%20-%20@%201139.68&HorizontalLineLeft5=%23color%3Dgreen%20Sev3%20%28p90%29%20-%20@%202616.33&HorizontalLineLeft6=%23color%3Dmagenta%20Sev3%20%28p99%29%20-%20@%206151.32&StartTime1=-P470D&EndTime1=-PT0H) | See how the pryon latency noticeably dropped after the sub-8-bit model’s deployment at the end of Aug, 2022. |