# 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 the runtime latency without impacting accuracy directly benefits the customers’ experience, which is critical for Alexa to increase the market share against the competitors. The goal of this work is to develop sub-8-bit quantization to compress ACE chip enabled on-device ASR (Laser/Theia) in order to reduce the Pryon engine latency (P50/90/99) without degrading accuracy, which is directly tracked in or related to ([Kingpin Goal](https://kingpin.amazon.com/#/items/257984) [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)** design and implement a computation paradigm for sub-8-bit quantization aware training which learns model parameters in a compressed, quantized state; **(b)** coordinate with ACE team to develop the hardware SDK to leverage the quantized model for acceleration; **(c)** release the first en-US Laser/Theia ASR model in sub-5-bit; **(d)** onboard new hires to adopt this tech innovation into more locales in multiple NNA enabled EFD Edge programs; and **(e)** coordinate 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 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 testsets 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)** delivers 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 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. Understands 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, make them auditable, available, and accessible. * You take a long term view of the business objectives, system-wide view of 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, AHS-ASR team has enabled 8-bit quantization aware training (QAT) in AHE-Lite (kingpin A-team goal) which was an important program for the launch of Raven, the Fire TV Cube 2nd gen device. However, the seemingly natural extension of going from 8-bit to sub-8-bit (5-bit) quantization was actually considered rather challenging. Admittedly, using 5-bit quantization will reduce the NNA bandwidth by 37.5%, which effectively reduces the user-perceived latency as a critical metric for Alexa’s customer experience. Nonetheless, going from 8-bit (corresponding to 2\*\*8=256 distinct weight values) to 5-bit (only 2\*\*5=32 weight values) heavily limits the dynamics of model weights, and hence its capacity, posing a major concern on the accuracy side. Furthermore, unlike 8-bit quantization where 255 quantization centroids are linearly fixed in a closed interval, 5-bit quantization requires finding the optimal locations for 32 centroids to make accuracy at a viable level: simply adopting the previous 8-bit QAT method to the 5-bit case would explode up the accuracy by over 10%. There has been a constant debating on how to effectively compress on-device ASR models to 5-bit via a highly reusable, end-to-end solution. In a nutshell, sub-8-bit quantization is pervasively considered an intrinsically **ambiguous** and **complex** optimization problem.  Secondly, the level of **Ambiguity** becomes even higher when considering ASR model quantization as a problem of 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 running on NNA, will 5-bit model be executed faster? In this project, I sent the feature request intake tickets from the early stage to ACE team, so as to have hypothesis verified in time and invest on promising co-design directions. Tacking the **ambiguity** with this mechanism, we observed the frame processing rates for 5-bit baseline to be 17.1% faster than 8-bit-compressed models. In parallel, we proposed to build Lloyd-Max algorithm in 5-bit quantization to optimally locate 32 quantized weight centroids. As a consequence, at runtime ACE team enables post-training quantization via K-means clustering algorithm accordingly (matching our in-training Lloyd-Max algorithm) to process our 5-bit ASR models. The outcome of this successful Software/Hardware Co-Design is that compressing models to 5-bit gives zero accuracy degradation.  Last but not the least, the available guidance, when we pushed our on-device ASR quantization solution to the cloud scenarios, is rather limited. One fundamental reason is again the intrinsic **ambiguity** of the project, leaving numerous questions unclarified before. For example, has cloud Conformer already been executed in 8-bit on 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? Yet, I managed to bond with partner teams and release owners from cloud ASR persistently to align on all those questions. Particularly, in partnership with NeMoRT (Chris Beauchene, Sr. SDE), we spotted the discrepancy between how the model is trained in 32-bit and how it’s executed on CPU in 8-bit, and explained why the discrepancy leads to quantization loss that hinder the cloud ASR model accuracy in WBR test sets. From there, we collaborated with ASR-EU team (Jahn Heyman, Sr. Applied Scientist) to test GQ out on every single stage from core-transducer training, checkpoint averaging, incremental learning, to neural biasing. With tremendous efforts, I was able to deploy GQ to cloud de-DE ASR with 1-3% WERR on glidepath and tail, which is recovered from the previous quantization induced loss. Later in that year, GQ was productized in en-GB and en-AU. Even with limited guidance, I took the ownership, bore with obstacles and production pitfalls, successfully navigate through the **ambiguous** process, and eventually completed the project in Runtime Modeling program.  **Scope of Influence:**  Making influence on other teams is never easy especially for an L5 applied scientist. The challenge becomes quite noticeable when I bring on-device tech innovations to the cloud teams, as usually the influence flow is the other way around. To address this challenge, I proactively scheduled 1:1 meetings and even created questionnaires collecting inputs and understanding concerns from cloud ASR teams (Hitesh Tulsiani, Sr. Applied Scientist in ASR-BLR, Jahn Heyman, Sr. Applied Scientist in ASR-EU, Harish Arsikere, Pr. Applied Scientist in ASR-BLR, etc) and NeMoRT team (Chris Beauchene, Sr. SDE). Working backwards from their suggestions and even doubts, I proposed tentative solutions with detailed figures and numbers for another round of discussions. Throughout this iterative process involving software/hardware experts from multiple teams, we gradually built the consensus that incorporating 8-bit GQ to cloud Conformer core-transducer training is an accomplishable goal in Runtime program and with great importance to improve the model accuracy on Tail [Kingpin].  For on-device ASR (Bluebottle/Crosstown), to ensure the timely release, I need to not only align with ACE team (Raviteja Chinta, Sr. SDE) on co-designing the 5-bit NNA SDK in time, but also **actively mentor and coach the scientists** working directly in productizing 5-bit quantization. For example, I onboarded Yi Xie (Applied Scientist II in AHS-ASR team) to productizing 5-bit quantization in en-US Crosstown model; I also created Runbooks and recorded a video tutorial for ODIE/Crosstown TVM packaging to mentor Rohit Barnwal (Applied Scientist I in AHS-ASR team, now at Tiktok) for en-GB Crosstown model release. Consequently, our team were able to incorporate 5-bit quantization to 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 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, influencing multiple teams with aligned consensus on the roadmap, rather being the consequence, is the premise for me to push 5-bit quantization to on-device ASR in Bluebottle/Crosstown, and 8-bit General Quantization to cloud ASR. I also actively present our research and development progress on 5-bit quantization in multiple internal venues, such as wake-word team meeting, core-transducer workstream LR, AMLC workshop, etc.  **Scientific and Technical Complexity:**  Sub-8-bit quantization were rarely investigated as they require sub-8-bit operators on neural network accelerators (NNAs), which often have inferior performance compared to their 8-bit counterpart 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 aggregates 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 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 the 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, called General Quantizer, or GQ.  Inspired from 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-n-play solution ASR quantization.  **Impact:** Sub-8-bit/general quantization has made board 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. Dr. Zhen 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), Dr. Zhen is 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). Dr. Zhen also onboarded a few other applied scientists to productizing 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, Dr. Zhen kept pushing the envelope and brought more innovation to ASR model in-training quantization, called General Quantizer (GQ), making the recipe more concise, 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. Dr. Zhen 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 is productized for cloud Conformer in de-DE, en-GB and en-AU locales.  ***Research impact:*** In collaboration with hardware experts from ACE team, under the guidance of senior and principal scientists, Dr. Zhen has 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*   Dr. Zhen has 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:**  Though being an applied scientist, I value the importance of intra-team and cross-team coordination just as much, especially during the project design/scoping stage, which helps building up consensus early on and consequently leads to highly reusable end-to-end solutions. For example, I took the initiative to communicate with Jahn Heymann (Sr. Applied Scientist in ASR-EU) and Chris Beauchene (Sr. SDE in NeMoRT) to understand the status-quo of the comprehensive release procedure of Conformer based cloud ASR and how Conformer is executed on CPU using INT8Hybrid GEMM. By harmonizing different views/suggestions between scientists and engineers, I came up with solutions that was consistent and well-integrated with works being done by AHE Engine team. As a consequence, our published science work on “General Quantization (GQ)” was successfully deployed in de-DE cloud Conformer on time, leveraging NeMoRT’s INT8 kernel for runtime acceleration yet with 1-3% WERR on WBR/Tail test sets thanks to reduced post-training quantization-induced loss. Later, it was deployed to en-GB and en-AU locales for cloud ASR by Venkata Kishore Nandury (Sr. Applied Scientist in ASR-BLR) with minimal support from me, which is a testimony of GQ’s reusability as an end-to-end 8-bit quantization solution.  For compressing on-device RNN-T based ASR models, it was intensively tracked as one of the top priority program of AHS-ASR (Bluebottle and Crosstown). Still, it necessitates a significant amount of cross-team joint effort, especially among AHS-ASR, ACE and Automation, as coordinated by Hieu Duy Nguyen (AS manager in AHS-ASR) and Brian Collins (Sr. TPM in AHS-ASR). I intensively collaborated with Raviteja Chinta (Sr. SDE at ACE, now at Meta) to have 5-bit prototypes tested on NNA v1 and v2 with verified latency and memory benefits, and have the 5-bit weight conversion automated with the automation team. I also took the initiate to come up with the runbook of converting NNAv1 model to NNA-v2 compatible version. Because of all these execution endeavors, we have deployed 5-bit quantization enabled ASR architectures to all of our NNA-enabled EFD Edge programs (Bluebottle and Crosstown) across all locales in 2022. 5-bit quantization reduced the size of our Echo/Echo-dot models by over 30% without losing accuracy: with that, Alexa is able to reach more customers from more device types with more constrained memory size, such as Cannoli/CheeseCake, the 5th generation of Echo Dot.  **Knowledge:**  As described above, I’ve demonstrated a deep and broad set of skills in the domain of ASR model compression during the procedure of achieving goals for RNN-T based Bluebottle/Crosstown programs and Conformer-based cloud ASR runtime program. Our science works for “sub-8-bit” quantization is highly recognizable as the TOP search on Google drawing citations from prestigious research and industrial institutes. Productizing any of these tech innovations requires way more than grasping the specialized expertise itself but a deep understanding of most relevant ASR components such as acoustic modeling, language modeling and end-pointing, to form a feasible solution. My knowledge base in both model compression and Alexa ASR’s tech stack allows me to conduct quick prototypes and experimental validation from unit-level in-training (Phasa/PyRama) to service-level end-to-end testings on accuracy (Djinn and DoryBlueshift-ASR) and latency (Physical-Device-Benchmarking-Portal), which facilitates several production deployments for both on-device and cloud ASR scenarios bringing significant benefits to Alexa’s world-wide customers, especially in the era of Generative AI with Large-ASR and NextGen-ASR.  **General qualifications:**  I was firstly hired by AHS-ASR in May 2020 as an applied scientist intern whose work was selected as one of the 17 best poster presentations (under the supervision of Hieu Duy Nguyen and Athanasios Mouchtaris). Later I joined AHS ASR group at Amazon PIT full time, driving several neural efficiency projects ever since for both on-device and cloud ASR with a strong production/research impact. Prior to Amazon, I worked on BERT-based multi-modal recommendation systems as a machine learning relevance intern at LinkedIn in 2018 and 2019. I received B.S. degree in Software Engineering from Xidian University in 2012 and 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 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. I’m on 6 US patents as an inventor. |

|  |
| --- |
| **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, lower-bandwidth, and of higher predictive performance. For Bluebottle program, corresponding to Echo, Echo Show and FireTV products, reducing the accuracy gap against 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.  **“Pushing the needle” of model size and user-perceived latency**  For 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 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 12.52% p50 UPL reduction and 6.31% WERR on glidepath accuracy.  **Unifying quantization tooling between edge and cloud ASR computing**  In Runtime Modeling program, the next generation of S8BQAT, General Quantizer (GQ) was integrated in 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 goes 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 to flagship speech processing international conferences for publications (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 from the Llyod-Max compression theory with a practical adaptation for a feasible computational overhead during training. I demonstrate influence over multiple teams (ACE team, ASR-EU team, NeMoRT team, etc) via a collaborative effort to bring (sub)-8-bit training to both on-device and cloud ASR for our customers. I also took the ownership to bring the scientific impact outside Amazon: throughout this project, I have published 2 peer-reviewed papers and 1 patent application, as the first author or inventor, based on his work on quantizing with co-authors from multiple organizations.  The science and production impact demonstrated though this project is a testimony of my readiness of becoming a senior applied scientist, bringing more novel methods and innovations into a range of customer products with fruitful commercial value and real-world applicability.  **Invent and Simplify:** I took a long-term view, constantly seeking ways to optimize and simplify the existing neural efficiency toolset as Alexa ASR is transitioning from RNN-T to Conformer-based core-transducer. Because of his endeavor, PIT ASR team’s neural efficiency solution has been easy-to-use, and well adapted to the progress of the ASR model architecture over years. Even till this day, it’s considered as an inseparable ingredient for various ASR programs, such as AutoS2I, Large-ASR, and Runtime Modeling, etc.  **Bias for Action:** I was bias for action. I stayed contact with different teams to gather their insights. I constantly explore new ideas that might solve the pain point for technology integration towards customers’ benefit. I managed to complete the General Quantizer (GQ) project in Q2/2023, by bringing 8-bit quantization to cloud ASR. Note that GQ was among just a few projects completed in Q2, as the org was changing gear towards Large ASR modeling. This proves my capacity of taking calculated risk and reconciling conflicts among teams to deliver results in time, which is of great value in the often highly complex and ambiguous product development cycles.  **Ownership:** being an applied scientist, I’m never shy from making efforts on non-science related tasks, such as cross-team coordination, reconciling conflict of opinions among scientists and engineers, communicating on the production triage with TPMs, etc. I fully comprehend the necessity of owning every single step in the release procedure in order to put the technology innovations in production or at the customers’ hands. In particular, even after our technology went in production, I still thought on behalf of the company to initialize the patent filing process, so as to well protect Amazon’s IP from other competitors.  **Dive Deep**: ~~although I’m hired as an L5 applied scientist~~, I still frequently “stay connected with” multiple layers in the Alexa ASR’s tech stack, sometimes even cross-org teams, to foster a better understanding of Alexa’s ecosystem. I stay curious and skeptical about benchmarking results, even if they are well received by the audience. For example, when I productized GQ in cloud ASR, 8-bit HybridGEMM was already introduced by NeMoRT way back. People rarely challenged the fact that post-training quantization can be lossy, even via the dynamic quantization approach. Yet, I kept digging into the numbers with Sr. SDEs from NeMoRT and Sr. Applied Scientists from ASR-EU team, and eventually spotted the scenarios when the quantization-induced loss can hinder the accuracy performance, especially when the weight outliers exist. Because of the deep dive, I justified the business value of enabling GQ for cloud ASR, resulting in improved accuracy on Tail and WBR by leveraging the existing NeMoRT INT8 toolset. |

|  |  |  |
| --- | --- | --- |
| **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. |