**Tech Promotion Work Summary**

# Employee Instructions

1. **Partner with your manager to define a next-level goal** (e.g., project, research, presentation, etc.) that allows you to demonstrate one or more next-level capabilities, development areas, and/or Leadership Principles. Next-level criteria can be found in your [Role Guideline](https://ivy-help-center.talent.a2z.com/article/article-1568200618517-Gtmgg1nif) in the “Moving to…” a level section. If your guideline is missing this section, your manager will need to clarify which expectations and technical skills at the next level they want you to demonstrate.
2. **Document *Planned Work* in Part I**. The purpose of documenting a planned goal is to get clear on what is expected and how you and your manager foresee that it gives you the opportunity to demonstrate the next level. If you get a new manager, this can inform them of previous agreements and inform them about what you are working on (relative to the next level).
3. **When you complete the work, fill out Part II*.*** This is a best practice. If you change teams or get a new manager, this gives the new manager a documented record of the work you have done to demonstrate the next level. It is also a record of the impact you had and how your work demonstrates our [Leadership Principles](https://inside.amazon.com/en/About/corevalues_EN/Pages/LeadershipPrinciples.aspx).
4. **Provide at least two work samples that best demonstrate the work** **delivered** (e.g., scope, impact, expertise, technical skill, etc.). Work samples vary by job role. Examples:
   1. Technical work samples - Demonstrate technical and/or scientific expertise. These include scientific research papers, models, algorithm design documents, technical specifications, working mockups/prototypes, code samples, code review commentary (both given and received), architecture/design documents, patent submissions, etc.
   2. Standard work samples - Demonstrate other skills. These include PR/FAQs, narratives, Product Roadmaps, OP1 documents, MBR/QBR materials, functional specifications, UX designs, presentation decks, broadcast videos, project proposals/ROI case studies, wikis, resource planning documents, etc.
5. **After your manager has documented their feedback in Part III, discuss whether it would be helpful to gather additional peer and/or customer feedback**. This is highly recommended, but not required. Your manager gathers the feedback and decides whether some comments are appropriate to add to **Part III**. Work Summary feedback providers can be any level, however higher levels offer better next-level perspectives.

# Manager Instructions

1. **Partner with your employee on a one or more goals that give them the opportunity to demonstrate the next level**. Review the next-level criteria defined in the employee’s [Role Guideline](https://ivy-help-center.talent.a2z.com/article/article-1568200618517-Gtmgg1nif). Design projects or other goals that deliver value for the team’s customer, technology, and business domain. Goals can also give them the opportunity to demonstrate Leadership Principles or resolve a gap identified when hiring or in a previous attempt to promote**.**
2. **If the employee is getting a Planned or Completed Work Review** – Make sure the employee’s summary is clear and well written.
3. **When the work is completed:**
   1. Evaluate the quality of the employee’s work and the results they delivered against the next-level capabilities (and Tech bar) expressed in the employee’s Role Guideline
   2. Document your feedback in **Part III**
   3. Make sure their work samples demonstrate key skills required at the next level
   4. Discuss whether there are peers or customers the employee thinks would be good to get feedback from about the specific effort in the work summary.

# PART I: PLANNED WORK

(Part I required for a [Planned Work Review](https://ivy-help-center.talent.a2z.com/article/article-1588633382719-aPKCqTQob).)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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)** | | | | |
| *Use your Role Guideline 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, which is pivotal factor for Alexa's market competitiveness. This project aims to develop a unified arbitrary bit-depth weight quantization solution for both cloud and on-device ASR models that facilitates the improved accuracy and runtime efficiency for cloud Conformer based ASR, and reduced memory footprint and user-perceived latency for ACE chip-enabled on-device ASR products. For cloud ASR, the primary goal is to enable 8-bit GQ in training to improve the accuracy in Tail and WBR by limiting the quantization-induced loss (Kingpin Goal [635862](https://kingpin.amazon.com/#/items/635862)), while leveraging the established NeMoRT’s dynamic quantization GEMM to realize the INT8 runtime. Regarding on-device ASR, the primary goal is to deploy 5-bit GQ to decrease Pryon engine latency (P50/90/99) while preserving accuracy, directly linked to Kingpin Goals ([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 General Quantizer (GQ), which learns model parameters in a compressed, quantized state for any bit-depth settings; (b) aligning with Alexa EU team to deploy 8-bit GQ in cloud de-DE Conformer based ASR model, and then Alexa Bangalore team for enabling 8-bit GQ in en-AU and en-GB locales; (c) coordinating with ACE team to develop the hardware SDK to leverage the 5-bit GQ quantized model for acceleration; (d) releasing the first en-US Laser/Theia ASR model in 5-bit; (e) mentoring new hires to adopt this tech innovation into more locales in multiple NNA enabled EFD Edge programs. The goals are measurable in the following way: (a) the accuracy from Tail and WBR test sets for cloud ASR should be improved by incorporating 8-bit GQ in the training process; (b) 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; (c) to achieve 15% relative latency reduction measured via Pryon engine latency (P50/90/99); and (d) the accuracy degradation in terms of WER on QBR test sets when enabling 5-bit GQ for on-device ASR (Bluebottle/Crosstown) should be less than 1.5% relative. To make the goal actionable, I 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) is able to design novel algorithms on existing neural efficiency toolsets. To keep the goal realistic, in terms of implementation and timeline, I need 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, I need 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.” * **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. * **Hire and Develop the Best:** Leaders raise the performance bar with every hire and promotion. They recognize exceptional talent, and willingly move them throughout the organization. Leaders develop leaders and take seriously their role in coaching others. We work on behalf of our people to invent mechanisms for development like Career Choice. | | | | |
|  | | | | |

# PART II: Results

(Complete this section after you finish the work. Parts I, II, and III required for [Completed Work Reviews](https://ivy-help-center.talent.a2z.com/article/article-1588633382719-aPKCqTQob).)

|  |
| --- |
| **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) for various Bluebottle and Crosstown devices. From there, we aim for further benefiting Alexa’s customers by (a) improving the cloud model’s predictive performance by incorporating 8-bit QAT to cloud ASR’s existing post-training 8-bit runtime, (b) going from 8-bit to 5-bit to further reduce the model size and user-perceived latency (UPL) for Bluebottle and Crosstown models. However, each of these goals posed unexpected challenges due to considerable ambiguity.  To begin with, it’s unclear if the existing cloud 8-bit post-training quantization (PTQ) can adequately preserve accuracy. While 8-bit PTQ has generally demonstrated robustness and has supported 8-bit cloud ASR runtime for years, the 8-bit dynamic PTQ can be lossy, a factor overlooked in cloud Conformer training. Therefore, the debate revolves around whether or not to enable in-training quantization to compensate the PTQ-induced loss.  Additionally, the previous 8-bit QAT method is constrained by multiple necessary calibration and analysis steps, making it challenging to seamlessly scale to cloud Conformer-based ASR models and 5-bit on-device ASR models. For instance, scientists typically spend an average of 2 days adopting the conventional 8-bit QAT approach. Therefore, it becomes imperative to propose a unified, model-agnostic, plug-and-play quantization scheme applicable to both on-device and cloud ASR models, regardless of whether they are RNN-T or Conformer-based.  Moreover, while employing 5-bit GQ reduces NNA bandwidth by 37.5% — significantly enhancing user-perceived latency which is crucial for Alexa’s customer experience — it heavily restricts model weight dynamics and capacity, raising accuracy concerns. A naïve 5-bit QAT implementation leads to 10% relative accuracy degradation. Therefore, the debate persists on designing a General Quantizer (GQ) supporting arbitrary bit-depth settings for both cloud and on-device ASR — a complex and ambiguous optimization challenge.  Lasts but not least, in the realm of ASR model quantization, the complexity rises, particularly in cross-team Software/Hardware Co-Design: it makes no difference to performance if the underlying hardware can’t take advantage of quantized models. For on-device ASR utilizing NNA, the critical question arises: will the 5-bit model execute faster? I coordinated with ACE team early on 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 explored the literature and invented General Quantizer (GQ), inspired from gradual sparsification and our internal sub-8-bit QAT technique. This successful Software/Hardware Co-Design venture yielded zero accuracy degradation when compressing models to 5-bit. It also significantly reduced scientists' adoption time for our quantization approach by 75%, from an average of 2 days to 4 hours.  **Scope of Influence:**  I was able to influence multiple teams to push General Quantizer into multiple locales for both on-device and Cloud models.  For instance, I took the initiative to schedule bi-weekly 1:1 meetings and created questionnaires to gather insights and address concerns from several key individuals, including Hitesh Tulsiani (Sr. Applied Scientist from ASR-Bangalore team), Jahn Heymann (Sr. Applied Scientist from ASR-EU team), Harish Arsikere (Pr. Applied Scientist from ASR-Bangalore team), and Chris Beauchene (Sr. SDE from NeMoRT team), among others. Engaging with cross-team members in this manner significantly expanded the reach of our model compression techniques. This expanded engagement led to discussions highlighting how our innovation could significantly benefit the accuracy of cloud ASR. It also sparked conversations on how our innovation could be further improved to be better incorporated by the release owner. 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 on WBR and Tail test sets by 2-3% relative.  In addition, 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 GQ 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) for the release of the en-GB Crosstown model. As a result, our team successfully integrated 5-bit GQ across all locales for Brownie/Ganache and Cannoli/CheeseCake. With 5-bit GQ, 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 GQ achieved 373.00 msec from 788.00 msec, or over 50% user-perceived latency reduction, which is a huge win for customers’ experience. Moreover, I collaborated closely with the ACE team (Raviteja Chinta, Sr. SDE in ACE) to co-design the 5-bit NNA SDK. And I actively showcase our research and development progress on General Quantizer across various internal forums, including the wake-word team meeting, core-transducer workstream LR, and the AMLC workshop.  **Scientific and Technical Complexity:**  Designing a general-purposed quantization-aware training scheme (General Quantizer) that can apply seamlessly to both cloud and on-device ASR in arbitrary bit-depth settings has been rarely investigated in the field due to its scientific complexity and the disparities between cloud and on-device ASR models. These disparities encompass both the model architecture and the runtime deployment aspects.  Regarding model architecture, the core-transducer in cloud ASR has transitioned from RNN-T to Conformer, while on-device ASR continues to utilize RNN-T as the acoustic model. Notably, RNN-T comprises a relatively limited number of dozens of kernels, whereas Conformer can exceed a hundred kernels. Moreover, the weight distribution between the recurrent kernels in RNN-T and the convolution/attention kernels in Conformer can significantly differ.  In terms of runtime deployment, cloud ASR has already supported INT8 runtime for most kernels via NeMoRT’s post-training quantization at ONNX level during Dory-Blue-Shift packaging. During this dynamic quantization process, the min and max values for every output channel of each matrix are mapped to the min and max values of the integer range, respectively. INT8 post-training dynamic quantization has generally proven reasonably accurate. As a consequence, to push our 8-bit GQ to cloud ASR demands in-depth analysis to validate several hypotheses: (1) Is the current INT8 post-training dynamic quantization genuinely lossless, and under what circumstances might loss occur? (2) Is it truly advantageous to enable quantization-aware optimization during training to account for the weight precision loss at test time? (3) How can we design the General Quantizer to seamlessly accommodate various types of Conformer weights, such as feedforward dense kernels, convolution kernels, and multi-head attention kernels, in a plug-and-play manner so as to be eventually adopted by the cloud ASR model release owner?  In addition to the complexity from cloud ASR’s perspective, the complexity of devising General Quantizer is compounded by the challenges posed in on-device ASR where the bit-depth goes from 8 to 5. (1) In contrast to 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%. (2) While our previous sub-8-bit QAT approach provided kernel-specific regularizations to determine optimal centroid locations, it necessitates model-pretraining, weight distribution analysis, and the incorporation of hand-crafted regularizers for all 5-bit kernels. This process proves time-consuming and hampers release owners' efforts to enhance customer experience through faster model execution using 5-bit runtime. In summary, creating a self-adaptable 5-bit General Quantizer that accommodates diverse weight distributions while also ensuring easy adoption by release owners for product integration proves fundamentally challenging.  **Impact:**  Sub-8-bit/general quantization has made a broad impact in terms of both research and production.  **For cloud ASR production:** Along with ASR-EU team, I deployed GQ in de-DE v59 - the first unified cloud Conformer model ([1] [Wiki page](https://wiki.labcollab.net/confluence/pages/viewpage.action?pageId=2054818930) [4] [Launch Announcement](https://quip-amazon.com/bTBQAjhiM2KM/Launch-Announcement-8-bit-Cloud-Conformer-Training-via-General-Quantization), [5] [LR Highlight](https://quip-amazon.com/nykpAZKcR9kV/Runtime-Modeling-Bi-Weekly-Update-07112023#temp:C:TAZ529cda86fbcf4f208965c3b3f)) that was trained with General Quantization (GQ), which yields 2% WERR on Tail (from 7.41% to 7.24%), 3% WERR on WBR (from 7.62% to 7.41%) (Kingpin Goal [635862](https://kingpin.amazon.com/#/items/635862)) and 8% WERR on Mshop\_eval\_live\_so\_latest (from 11.90% to 10.89%) (also highlighted in QBR of 2023/Q2). As an 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 dynamic 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 ASR in de-DE, en-GB and en-AU locales.  **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 ([2] [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, with 8.5% relative p50 UPL reduction. Similarly, for CrossTown (with devices of more constrained memory size), I am the release owner of the es-ES model ([3] [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 while reducing the slot-WER from 6.64% to 6.14% (by 7.53% relative) and p50 UPL by 12.52% (see [6] the [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)), which successfully facilitates deploying our ASR models to 5th generation of Echo Dot (Cannoli/CheeseCake).  GQ effectively lowers the endeavor of the release owner to intake our innovation by 2 days to 4 hours on average. I 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 5-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 Alexa.  **Research impact:** In collaboration with Raviteja Chinta (Sr. SDE at ACE), Tariq Afzal (Pr. ML Architect at HW Compute Group), Hieu Duy Nguyen (Manager at AHS-ASR), Anastasios Alexandridis (Sr. AS at AHS-ASR), Athanasios Mouchtaris (Sr. Manager at AHS-ASR), and Ariya Rastrow (Pr. AS at Alexa AI), I have secured the invention via a filed patent application [9].  I have also published academic papers for sub-8-bit quantization in top-tier conference proceedings, demonstrating Alexa’s leading role in the Automatic Speech Recognition (ASR) / Machine Learning (ML) field [7,8].  **Execution:**  To have a mechanism that works with both Cloud and on-device models, I put much effort into building consensus with Cloud model owners, NemoRT and ACE engineers, resulting in highly adaptable end-to-end ASR model compression 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, "General Quantization (GQ)" was effectively implemented in the de-DE cloud Conformer, leveraging NeMoRT’s INT8 kernel for runtime acceleration. We carefully designed the quantization mechanism for GQ to ensure it’s adaptable and straightforward for release owner to intake, and merge all required codes in Phasa mainline early on. This implementation notably resulted in a 2-3% improvement in WBR/Tail test sets due to reduced post-training quantization-induced loss. Because of the production deployment for de-DE and GQ’s highly adaptable plug-and-play algorithm design, 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 the technology owner.  I also collaborated closely with Raviteja Chinta (Sr. SDE at ACE) 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. Because of my persistent effort on coordination with other teams, 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, along with 12.52% p50 user-perceived latency (UPL) reduction. 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:**  Successfully productizing 5-bit quantization-aware training / general quantization to both on-device and cloud ASR systems demand more than just specialized expertise. It necessitates a comprehensive understanding of the full model release process, not just limited to acoustic model and language model trainings but also the integration to neural biasing and incremental learning, 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).  **General qualifications:**  I was first hired by AHS-ASR in May 2020 as an applied scientist intern, already publishing 1 paper during the internship. Later, I joined the AHS ASR group at Amazon PIT full-time, spearheaded several impactful neural efficiency projects for both on-device and cloud ASR, leaving a significant mark on production and research initiatives. 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. 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 M.S. in Computer Science from Tsinghua University in 2015 and a B.S. degree in Software Engineering from Xidian University in 2012 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. |

|  |
| --- |
| **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?* |
| In this project, we unified quantization tooling between edge and cloud ASR computing via General Quantizer (GQ). In the Runtime Modeling program, the next generation of S8BQAT, 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.  General Quantizer (GQ), as a Software Hardware Co-Design solution for on-device speech recognition, makes Alexa faster, has lower bandwidth, and higher predictive performance. For the Bluebottle program, corresponding to Echo, Echo Show, and FireTV as our flagship on-device products, reducing the accuracy gap against the cloud model is of high priority (Kingpin Goal [635862](https://kingpin.amazon.com/#/items/635862)) to achieve customers’ satisfaction. As a concrete example, GQ 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 being larger, the model achieves 8.56% p50 UPL reduction (from 1391ms to 1272ms). 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. GQ also elevated 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. By the end of 2022, we have deployed 5-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 Alexa.  In additional to the production impact, we were also able to influence the research community with two papers [7, 8] published in prestigious conferences (Interspeech’22 and IEEE SLT’23, respectively). Furthermore, to secure the intellectual property, we also filed a patent application [9]. |
|  |

|  |
| --- |
| **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 envisioned the necessity of achieving a unified design of varying bit-depth quantization-aware training tool set for Alexa by considering both cloud ASR and on-device ASR scenarios altogether. Throughout cross-team collaboration and rigorous execution, we proposed General Quantizer, with scientific impact of filed patent [9] and publications [7, 8], productized in both cloud and on-device ASR bringing our Alexa’s innovations to millions of customers.  **Dive Deep**: Despite the prevalence of post-training dynamic quantization in cloud ASR, I delved into details, collaborating with NeMoRT's Sr. SDEs and ASR-EU's Sr. Applied Scientists. This scrutiny unveiled scenarios where quantization-induced loss impacted accuracy, particularly in the emergence of weight outliers — a common occurrence, particularly in larger models. This in-depth investigation led me to validate the potential of implementing GQ for cloud ASR, ultimately resulting in enhanced accuracy for 2% WERR on Tail (from 7.41% to 7.24%) and 3% WERR on WBR (from 7.62% to 7.41%) (Kingpin Goal [635862](https://kingpin.amazon.com/#/items/635862)) using NeMoRT's established INT8 HybridGEMM toolset. This underscores the tangible benefits arising from being self-critical and diving-deep regarding seemingly established metrics.  **Bias for Action:** I operated with a bias for action, maintaining continuous communication with various teams to gather diverse insights. Sometimes, I even work beyond orders so as to constantly explore innovative ideas aimed at addressing technological integration challenges for the benefit of our customers. For example, even though our previous 8-bit QAT has been well established and going towards 5-bit seems a natural extension, I insisted on a more general, self-adaptable quantization-aware training paradigm which can be easier to adopt for production deployment, and better scale up to cloud ASR at the same time. My “bias for action” noticeably reduces the technology intake effort: what would take the release owner 2 days to roll out 5-bit quantization, can now only take 4 hours. This hugely facilitate the on-time ASR model release for Bluebottle and Crosstown.  **Ownership:** I actively maintain connections across partner teams within the Alexa ASR org and even engage with cross-organizational teams. This proactive engagement is aimed at cultivating a comprehensive understanding of Alexa's ecosystem. More importantly, it facilitates productizing our technology innovation on sub-8-bit / General Quantization. For example, I coordinated with Hitesh Tulsiani, Jahn Heymann, Harish Arsikere and Chris Beauchene, to build consensus on the necessity of incorporating in-training quantization in the early stage. In collaboration with ASR-EU team, I rigorously conducted experimental validation from not just in the core-transducer training stage, but the incremental learning phase on meta dory, along with neural biasing to seek alignment on GQ’s accuracy benefit iteratively from each model release stage till the very end. Without such level of commitment as an owner, GQ wouldn’t bring benefits to our customers. I also drove the IP submission process: I drafted the document for provisional application, answered layers’ questions in several rounds of discussions, coordinated with Raviteja Chinta (Sr. SDE at ACE) and Tariq Afzal (Pr. ML Architect at HW Compute Group) to strengthen the innovation from the hardware’s aspect, all of which eventually result in the successful patent filing.  **Hire and Develop the Best:** throughput this project, I mentored/coached many applied scientists in the group, developing them to become acquainted in using our sub-8-bit /General Quantization to bring accuracy, latency benefits to our customers. For example, I played a key role in guiding Yi Xie (Applied Scientist II, AHS-ASR team) through the implementation of 5-bit quantization for the en-US Crosstown model. Additionally, I mentored Rohit Barnwal (Applied Scientist I, AHS-ASR team) for the launch of the en-GB Crosstown model. Consequently, our team successfully integrated 5-bit quantization across all local variations for Brownie/Ganache and Cannoli/CheeseCake. The adoption of 5-bit quantization led to a remarkable reduction of over 30% in the size of our Crosstown models. This effectively addressed the business challenge caused by memory constraints on NNA-v2 for Cannoli/CheeseCake. Specifically, for Stage1 Pryon Latency (ms) on NNA v1, the implementation of 5-bit quantization resulted in reducing latency from 788.00 msec to 373.00 msec—a user-perceived latency reduction of over 50%. This significant improvement noticeably enhanced our customers’ experience. |

|  |  |  |
| --- | --- | --- |
| **WORK SAMPLES** | | |
|  | **Work Samples Name and Link** |
|  | [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) | |
|  | [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 | |
|  | [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. | |
|  | [Sub-8-bit quantization-aware training](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) | Conference: Interspeech, 2022  Title: Sub-8-Bit Quantization Aware Training for 8-Bit Neural Network Accelerator with On-Device Speech Recognition  Authors: **Kai Zhen**, Hieu Duy Nguyen, Raviteja Chinta, Nathan Susanj, Athanasios Mouchtaris, Tariq Afzal, and Ariya Rastrow  **Deployed to on-device ASR.** | |
|  | [General quantization](https://assets.amazon.science/0c/03/41fc077547799c2350ccb3a4ac15/sub-8-bit-quantization-for-on-device-speech-recognition-a-regularization-free-approach.pdf) | Conference: Spoken Language Technology Workshop (IEEE SLT), 2023  Title: Sub-8-Bit Quantization for On-Device Speech Recognition: A Regularization-Free Approach  Authors: **Kai Zhen**, Martin Radfar, Hieu Nguyen, Grant Strimel, Nathan Susanj, Athanasios Mouchtaris  **Deployed to both on-device and cloud ASR.** | |
|  | Patent on ASR model quantization | File Number: P77898-US01  Title: COMPRESSION OF MACHINE LEARNED MODELS  Inventors: **Kai Zhen**, Hieu Nguyen, Raviteja Chinta, Tariq Afzal, Anastasios Alexandridis, Athanasios Mouchtaris, Ariya Rastrow | |

**--SUMMARY SHOULD NOT EXCEED 3 PAGES--**

# PART III: Evaluation

|  |  |  |  |
| --- | --- | --- | --- |
| **Manager FeedbACK** | | | |
| *Date:* | | | |
|  | | | |
|  | | | |
| **Additional FeedbACK** | | | |
| **feedback provider Name:** |  | **Feedback provider business title:** |  |
|  |  |  |  |
| **Feedback date:** |  | **Feedback provider level:** |  |
| **Feedback type:** | Select | **Relationship to candidate:** | *(e.g., customer, peer, stakeholder, teammate)* |
|  | | | |
| **Additional FeedbACK** | | | |
| **feedback provider Name:** |  | **Feedback provider business title:** |  |
| **Feedback date:** |  | **Feedback provider level:** |  |
| **Feedback type:** | Select | **Relationship to candidate:** | *(e.g., customer, peer, stakeholder, teammate)* |
|  | | | |
|  | | | |