# 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.* | | | | |
| (WHAT IS THE GOAL OF THIS PROJECT? WHICH KINPIN GOAL DOES THIS ACHIVE BECAUSE OF THIS PROJECT?)  (DON’T MAKE IT TOO LONG)(READERS ARE NOT TECHNICAL ADAPT)  Representing ASR model’s weights in lower bit-depth (model quantization) for runtime efficiency has served a critical role for faster and lower-bandwidth speech processing, bringing Alexa to a wider audience in various scenarios, such as in in-car units and on portable devices, where Internet connectivity can be intermittent. From both technique and engineering’s perspectives, it is essentially challenging to heavily compress ASR models for the latency reduction while preserving the accuracy.  This challenge has been tackled in the Neural Efficiency project, as a cross-site joint effort among Alexa Hybrid Science (AHS) ASR team, ACE team, NeMoRT team, and engine team, etc. The neural efficiency project has also been closely integrated to not only AHS-ASR team for Bluebottle and CrossTown programs, but also Alexa Runtime Modeling program and Large-ASR program with broad product impact and cross-team endeavor.  As an L5 applied scientist aiming for the next level, Dr. Zhen’s particularly needs to fulfill the following job duties:   * Think big and insist on the highest standard to coordinate with the hardware team on further lowering the bit-depth from 8-bit to sub-8-bit for on-device ASR: this is critical to lower user-perceived latency without hindering accuracy to bring Alexa ASR to more device types and a wider audience. In particular, the newer Echo Dot devices are with more strict memory constraint which entails model compression for runtime deployment to be viable. Dr. Zhen needs to propose and implement sub-8-bit quantization, collaborate with ACE team on effectively accelerating 5-bit models via the SDKs, and apply the software-hardware co-designed compression paradigm to on-device ASR model release procedure for Alexa’s customers. * Invent and simplify advanced quantization mechanisms to unifying the neural efficiency recipe among on-device ASR models and cloud models: recipe unification is crucial to lower the cross-team coordination effort, workload of release owners, and above all, the cost. A reusable, model-agnostic, and plug-and-play design of our quantization method becomes a necessity to achieve the unified production intake. Dr. Zhen needs to collaborate with cloud ASR Aachen team and onboard to their model release process; adapts and modifies the existing quantization method to achieve generalizability; and finally, productizes the general quantization approach along with the tech lead from the Aachen team. * Push the envelope of the state-of-the-art quantization methodology: Dr. Zhen should publish academic papers in top-tier conference proceedings to demonstrate Alexa’s leading role in the ASR/ML field. If possible, Dr. Zhen should also secure 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. | | | | |
|  | | | | |
| **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.* | | | | |
| (COPY AND PASTE FROM THE GOAL GUIDLINE)  Next-Level Criteria:   * Has significant knowledge/expertise in one or multiple applied science disciplines * Is able to influence the technical (scientific and engineering) strategy of teams. Understands that not all problems are new (or require new algorithms) * Builds and owns ML solutions that are easy for others to contribute to. Knows how to document solutions, make them auditable, available, and accessible * Takes a long term view of the business objectives, system-wide view of product roadmap, technologies, and how they should evolve   (2 IS TOO LITTLE, ADD MORE!)  Leadership Principles:   * 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. * 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.* |
| (AMBIGUIITY)  BB – DON’T KNOW HOW TO COMPRESS TO 5-BIT, WHICH WAY? PTQ V.S. QAT? ACE OR SCIENCE? HOW TO PROCEEED  CLOUD – WHERE DOES THE DEGRAATION COME FROM. DEEP DIVE – PTQ, OUTLIERS  (SCOPE OF INFLUENCE)  MORE ABOUT INTERACTIONS TO BRING THINGS TO CLOUD/BB.  RAVI  JAHN  (SCIENTIFIC COMPLEXITY)  LIST THE DIFFICULTIES. HOW DIFFICULT IS IT. HOW DIFFULT TO IMPLEMENTATION? WHAT ARE THE DRAWBACKS?  (IMPACT)  BB/CLOUD, KINPIN (GET FROM THE LR)  ALREADY IN PRODUCTION  (EXECUTION)  HOW DID WE COORDINATE WITH MULTIPLE TEAMS TO ENSURE THAT THEY KNOW AND WOULD USE OUR TO LAUNCH ON TIME.  MECHANISMS.  (KNOWLEDGE)  KIND OF BLUR. EXPERIENCE ON COMPRESSING THE MODEL.  PHD – WORK, KNOW HOW TO COMPRESS THE MODEL.  (GENERAL QUALIFICATION)  COPY AND PASTE FROM CV.  In this project, Dr. Zhen addressed the need for a fast and accurate on-device speech recognition model interface by studying model quantization techniques. He 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. This study is significant because Dr. Zhen developed a model that is able to run much faster and save memory space (Exhibits 1-4, 8-22). Dr. Siegfried Kunzmann provides further detail regarding this research project (Exhibit 2).  Sub-8-bit quantization has made board impact, in terms of both research and production.   * It effectively lowers the user latency by over 30%, which affords a larger RNN-T architecture for improving recognition accuracy. Due to its clear implications for improving the performance and accessibility of voice recognition technology in electronic devices (Exhibit 29), Dr. Zhen’s development of a sub-8-bit quantization-aware training method has been utilized in multiple Amazon.com Inc. voice-controlled assistants, including the products Echo, Echo Dot, and Echo Show. Dr. Zhen is the release owner of the first sub-8-bit Bluebottle en-US model. * It reduces the memory-footprint by 30% for CrossTown ASR models on NNA v1/v2, which solved the memory bottleneck of deploying ASR models to 5th generation of Echo Dot (Cannoli/CheeseCake). Dr. Zhen is the release owner of CrossTown es-ES model. 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. * Conference proceedings: ***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.* * Patent: ***Kai Zhen****, Hieu Nguyen, Raviteja Chinta, Tariq Afzal, Anastasios Alexandridis, Athanasios Mouchtaris, Ariya Rastrow, Compression of Machine Learned Models, P77898-US01*   Insisting on the highest standards: 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 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. * 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, the 8-bit-trained de-DE model preserves 1-3% WERR on the test sets. GQ has been integrated in Bluebottle R18 en-US model for RNN-T based on-device ASR; besides, GQ is productized in 2023 for cloud Conformer in de-DE, en-GB and en-AU locales. * It has the foreseeable potential for INT4 quantization for RescoreBERT, UCE, etc. * Conference proceedings: **Kai Zhen**, Martin Radfar, Hieu Nguyen, Grant Strimel, Nathan Susanj, Athanasios Mouchtaris, "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. |

|  |
| --- |
| **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?* |
| (IMPACT/MEASURE OF SUCCESS)  BB/CLOUD, KINPIN (GET FROM THE LR)  ALREADY IN PRODUCTION  JUST REPHRASE IT AND ADD MORE DETAILS. WRITE IT FIRST AND SIMPLIFIES IT TO THE PREVIOS PART.  INSTANTIATE LPs.  Delievery  Dive deep  Bias for action.  In summary, throughout this project, Dr. Zhen consistently performs at the next level. The implementation of Dr. Zhen’s novel methods and innovations into a range of customer products reflects the commercial value and real-world applicability of Dr. Zhen’s work and further demonstrates that he has made substantial progress toward achieving his proposed endeavor of advancing current automatic speech recognition technology with novel machine learning techniques.  Scope of Influence: Dr. Zhen demonstrates influence over multiple teams (ACE team, Aachen ASR team, etc) via a collaborative effort to bring (sub)-8-bit training to both on-device and cloud ASR for our customers.   * Sub-8-bit quantization for on-device ASR: 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. * General quantizer (GQ) for cloud ASR: by the end of 2023, GQ has been incorporated in cloud Conformer for de-DE, en-GB and en-AU locales.   Science impact: Dr. Zhen has 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.   * Conference proceedings: ***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.* * Conference proceedings: **Kai Zhen**, Martin Radfar, Hieu Nguyen, Grant Strimel, Nathan Susanj, Athanasios Mouchtaris, "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. * Patent: ***Kai Zhen****, Hieu Nguyen, Raviteja Chinta, Tariq Afzal, Anastasios Alexandridis, Athanasios Mouchtaris, Ariya Rastrow, Compression of Machine Learned Models, P77898-US01* |
|  |

|  |
| --- |
| **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? |
| **(HOW TO HELP THE OTHERS?)**  **YI?**  **CLOUD RELEASE OWNERS?**  **PAPER/PATENT.**  **Invent and simplify:** Dr.Zhen 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-basesd 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:** Dr. Zhen was bias for action. He stayed contact with different teams to gather their insights. He constantly explores new ideas that might solve the pain point for technology integration towards customers’ benefit. He 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 can be considered as a testimony to his 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. |
|  |

|  |  |  |
| --- | --- | --- |
| **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 |