# 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.* | | | | |
| In the era of large ASR modeling with both Conformer and Rescorer succeeding 1B+ parameters, neural efficiency becomes critical to achieving “latency- and cost-neutral operating points”. Tracked in Large ASR Hardware Acceleration program, delivering the large hardware accelerated ASR model architecture into full production on GPU fleet entails a number of neural efficiency related innovations: along with existing neural efficiency techniques, such as INT4 Conformer [1, 2], weight sparsification is another way to save the number of bits for matrix transferring and computation.  This project focuses on one way to make neural networks run more efficiently — also a vital concern on resource-constrained devices — to reduce weights of small values to zero, A.K.A. large model sparsification. For example, consider a 2 by 4 matrix, [[*0.1*, 0.9, 0.8, -0.5], [*0.3, 0.2, -0.1*, -0.7]]. Sparsifying half of the weights based on their absolute values prunes the matrix to [[*0,* 0.9, 0.8, -0.5], [*0, 0, 0,* -0.7]]. Ideally, those “pruned-out” weights do not need to be transferred nor participate in matrix multiplication, and therefore the inference is faster. However, accelerating sparse ASR models requires hardware support: dense or sparse models would make no difference if the hardware could not take advantage of it.  To demonstrate at the next level Applied Scientist, Dr. Zhen needs to fulfill the following job responsibilities:   * **Be the owner**: being an Applied Scientist in the ASR org, Dr. Zhen needs to closely coordinate with experts from the hardware teams (such as the internal NeMoRT team and external team from Nvidia) in sync with the software-hardware co-design for sparsification, hardware readiness timeline, etc. Simply making efforts on the modeling side may not be sufficient to bring our innovations to the customers. * **Insist on the highest standards**: applying sparsity to ASR models without losing much model precision can be quite challenging. To achieve customers’ satisfaction, Dr. Zhen should aim at reducing 50% of the model weights to zero with less than 2% relative accuracy loss. To meet that goal, he needs to learn, develop, and improve Alexa’s policy-based sparsity-aware training strategy that enables the gradual reduction of low-value weights during training, so the network learns a model amenable to weight pruning. * **Frugality**: training large ASR models on P4 instances can be expensive and time consuming. Dr. Zhen needs to make innovations to just conduct model sparsification during finetuning for 1-2 days or even less, as opposed to training from scratch for 2 weeks on 16 P4 instances. This is crucial not only to reducing the training cost but also making the timeline viable for continuous integration of large ASR model release. * **Earn Trust:** deliveringhardware accelerated large ASR models with sparsity requires collaborations among multiple teams **early on.** Dr. Zhen needs to persistently update the partner teams and leaders, with a weekly/bi-weekly cadence, on not only the highlights but lowlights in a candid manner, such that he can gradually build consensus around the accuracy and latency impact along the journey. * **Have backbone; disagree and commit**: in attending the meetings for large-ASR LR and hardware-acceleration LR, Dr. Zhen shouldn’t be shy to voice his concern over the timeline and directions to pursue, etc. But once the decision is made, he needs to commit wholly so as to deliver all expected results in time. | | | | |
|  | | | | |
| **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.* | | | | |
| Next-Level Criteria:  Dr. Zhen consistently performs at the next level when executing this project, particularly in the following aspects:   * 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. * 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.   Leadership Principles:   * Deliver Results: Leaders focus on the key inputs for their business and deliver them with the right quality and in a timely fashion. Despite setbacks, they rise to the occasion and never settle. * Earn Trust: Leaders listen attentively, speak candidly, and treat others respectfully. They are vocally self-critical, even when doing so is awkward or embarrassing. Leaders do not believe their or their team's body odor smells of perfume. They benchmark themselves and their teams against the best. * 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.* |
|  |

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

|  |
| --- |
| **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 we influence other teams to adopt our innovation / mentor junior scientist / Repeat papers/deliverables here as well.**  **Hire and Develop the Best  Ownership** |

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