# 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. * **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. | | | | |
|  | | | | |

# 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:**  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 and Technical Complexity:**  LIST THE DIFFICULTIES. HOW DIFFICULT IS IT. HOW DIFFULT TO IMPLEMENTATION? WHAT ARE THE DRAWBACKS?  **Impact:** Sub-8-bit 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). 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:***  Dr. Zhen has 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, "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.   In collaboration with hardware experts from ACE team, under the guidance of senior and principal scientists, Dr. Zhen also 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*   **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 qualifications:**  COPY AND PASTE FROM CV. |

|  |
| --- |
| **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.­­ |
|  |

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