Deepwave Digital Creates an AI Enabled GPU Receiver for a Critical 5G Sensor

Whitepaper

Deepwave Digital: AI Enabled GPU Receiver for a Critical 5G Sensor

John D. Ferguson, Peter Witkowski, William Kirschner, Daniel Bryant

Deepwave Digital has leveraged the Artificial Intelligence Radio Transceiver (AIR-T) to create the first deep learning sensor for a 5G network. This network, the Citizens Broadband Radio Service (CBRS), will be the first spectrum sharing service provided by the telecommunications industry that leverages real time RF sensing. Critical to the operation of CBRS is the ability to determine if priority users are active on specific frequency channels. When no priority users are present, the spectrum may be reallocated for commercial networks to provide new enterprise services or additional bandwidth to existing services.

The AIR-T is a unique platform that combines radio frequency (RF) hardware with an embedded NVIDIA Jetson module for high throughput Digital Signal Processing (DSP). For CBRS, the Deepwave team has implemented a deep neural network (DNN) on the AIR-T that is capable of detecting, classifying, and reporting the presence of priority users with extreme accuracy. This work demonstrates that the GPU is a viable computational solution for both signal processing applications and real time deep learning inference.

This post begins by introducing dynamic spectrum allocation in telecommunications and then discusses the technology created by Deepwave Digital to incorporate deep learning into the field of signal processing. Additionally, real time GPU processing is discussed in terms of both traditional DSP approaches as well as deep learning methods. Finally, CBRS sensor, that leverages each of these technologies is discussed in detail.

1 Dynamic Spectrum Allocation

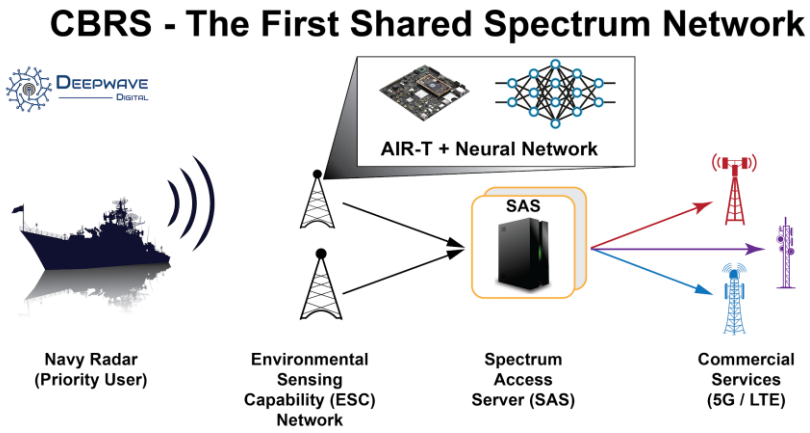

Figure 1: CBRS Network Operation

Communications systems function by transmitting and receiving radio signals between various nodes. These radio signals carry data content such as video, audio, or internet traffic. The recent explosion of IoT devices and LTE/5G enabled cell phones, spectrum congestion can degrade network performance and reliability. Historically, the spectrum has been managed by forcing each communication system to operate in a specific pre-defined, fixed frequency range. This system allows spectrum management to be simple but may result in large amounts of underutilized spectrum. For example, a block of frequencies may be allocated to a group of users who rarely utilize the spectrum, while another group of users may be stuck with less bandwidth than they require. It is often difficult to plan ahead and prioritize such use cases. A more advanced approach is to allow for dynamic spectrum allocation to maximize utilization and prioritize usage. This approach is typically referred to as spectrum sharing. While fully autonomous spectrum sharing is still a research topic, demonstrations involving the DARPA Spectrum Collaboration Challenge (SC2) have shown promising results.

In parallel to these research efforts, the first shared spectrum network will begin operation in 2020 as part of the United States' 5G rollout. This CBRS network will dynamically reallocate 100 MHz of spectrum in the 3.5 GHz band currently used for maritime radar and other purposes, but only at certain times and in certain locations.

As illustrated in Figure 1, the CBRS spectrum sharing process begins with an Environmental Sensing Capability (ESC) network monitoring a Dynamic Protection Area (DPA). The ESC determines if a priority user is operating (i.e., users such as radar who previously had sole authority to transmit on these frequencies nationwide). It then passes this determination to the Spectrum Access Server (SAS) which uses this information to repurpose the unused portions of the CBRS spectrum for commercial services like LTE. Any channels that the ESC determines are actively used by a priority user are blocked from any other usage during that time. While this spectrum sharing methodology may not be as sophisticated as some of the aforementioned research efforts, it is the first shared spectrum network that is viable for both commercial enterprises and government regulators and will pave the way for more advanced technologies and architectures to be fielded in the future. Moreover, such a design demonstrates what can be achieved with current radio and network technology and will serve as a model of how spectrum sharing can be done effectively.

This sensing capability, to determine if a priority user is active at a particular location, is the piece of new technology that is key to the CBRS network performing correctly. Not detecting a priority user's transmissions can result in significant degradation in the priority user's ability to properly function due to commercial interference. Conversely, falsely detecting transmissions (or misclassifying commercial users as priority users) will restrict other users and cause the spectrum to be underutilized.

In recent years, technologies have been developed to detect and classify features in images, signals, and other kinds of data. One such promising technology is the deep neural network. DNN classification algorithms have shown significant capability to process audio signals of similar structure with high accuracy for applications as varied as music recognition, speaker identification, earthquake detection, and gunfire localization. While DNNs are just beginning to be applied to radio frequency signals, the signal detection and classification problem posed by the CBRS network proved to be a prime candidate for this technology [1].

2 DEEPWAVE DIGITAL TECHNOLOGY

One of the common themes in modern RF systems is the tight coupling between the hardware that generates and receives radio signals and the software that controls this hardware and processes these signals. This is driven by the desire to minimize the time when the radio is not being used, miniaturize the system both in physical size and power required to operate it, and increase the rate that data can be received and sent. Realizing this, Deepwave has created an integrated hardware and software product to enable deep learning applications in RF systems: the Artificial Intelligence Radio Transceiver. Shown in Figure 2, the AIR-T is the first Software Defined Radio (SDR) designed specifically for deploying deep learning and GPU accelerated algorithms where the signal is found: at the edge. By embedding the processing necessary for DSP in the RF sensor, the AIR-T eliminates the usual headache of figuring out how to transfer large amounts of high bandwidth signal data back to a server.

Figure 2: Artificial Intelligence Radio Transceiver (AIR-T)

Historically, the hardware of choice for DSP applications has been the Field Programmable Gate Array (FPGA) due to the need for high throughput and low latency. Unfortunately, this high performance comes at the expense of long development times due to the specialized nature of the hardware. Software support is also often lacking and developers spend more time tinkering with the hardware and drivers than they do focusing on the actual algorithm. As CPUs became more capable, the SDR began to emerge, providing greater flexibility and ease of programming. In SDR, the mechanical components in traditional radio systems, such as filters and amplifiers, are replaced by software components in order to provide greater flexibility and reduced development time. This flexibility comes at the expense of reduced performance (increased latency and reduced throughput). Meanwhile, the trend with RF hardware (i.e, transceiver ASICs) has been greatly increased capability in terms of instantaneous bandwidth and the number of channels (for MIMO applications). In order to truly leverage these capabilities, modern RF engineers require a DSP system that can process the massive amount of data generated by new RF ASICs,with the agility and ease of use provided by the SDR.

In parallel to the advancements in SDR, there has been a trend to use Graphics Processing Units (GPUs) for general computing tasks. The GPU lends itself well to many DSP algorithms due to

the parallel nature of the underlying mathematical operations. However, not all operations are particularly well suited for GPU processing, and so a system using a traditional GPU may be forced to choose between either using the GPU inefficiently or copying data to and from the GPU multiple times, which also significantly limits the performance of the system. However, on the NVIDIA Jetson family of devices, the need for this operation is removed since the CPU and GPU share a pool of physical memory on the module itself. With this architecture, developers of DSP algorithms can treat buffers of signal data read from the RF frontend as GPU buffers and run their algorithms accordingly. This allows RF engineers to effectively use the processing power and flexibility that the GPU can provide while using the SDR programming model that they are already accustomed to. Furthermore, now that these signals exist in GPU memory, implementing a DNN to classify them is possible using existing frameworks and methods and without performance penalty. This design allows for the AIR-T to transmit and receive signals while simultaneously executing GPU accelerated signal processing and deep learning inference on a single, unified, platform.

Deepwave's AIR-T is equipped with an integrated software suite: AirStack. AirStack extends NVIDIA's Jetpack SDK with drivers, libraries, and applications that fully integrate the RF transceiver into the development environment and support any DSP application. Developers already familiar with the NVIDIA environment may leverage familiar libraries, APIs, and developer tools as they are included with AirStack.

3 Deep Learning as an CBRS Signal Classifier

A critical component of the CBRS network infrastructure is the ESC sensor network. Deepwave Digital has developed a DNN algorithm, as part of the ESC sensor, that has completed certification for deployment [2]. This sensor will be a key component of the service provided by Deepwave's strategic partner, Key Bridge Wireless. By leveraging the AIR-T and its AirStack development environment, Deepwave was able to implement a system capable of ingesting 125 MHz of bandwidth (4 Gbps) and determining whether or not a signal of interest is present. All the software necessary to receive, detect, classify, and make decisions about signals in the environment runs on a single NVIDIA Jetson TX2.

3.1 Signal Processing on GPUs

At GTC DC 2019, Deepwave's presentation outlined the various methods for performing DSP on an NVIDIA GPU and, in particular, the AIR-T. A key component of the AIR-T is the onboard Jetson TX2, which provides an ARM CPU and a Pascal GPU as computational resources. This section will go into depth on how to best leverage both these resources.

One of the most widely used SDR toolkits is GNU Radio. Like the majority of SDR applications, most functions in GNU Radio rely on a CPU or (with the addition of RFNoC) FPGA processing. Since many DSP engineers are already familiar with GNU Radio, Deepwave created two free and open source modules for leveraging GPU acceleration from within GNU Radio. The first one, GR-CUDA, provides a detailed tutorial on how to incorporate a CUDA kernel into GNU Radio. The second, GR-Wavelearner, is a framework for running both GPU-based FFTs and inference operations inside a GNU Radio application via cuFFT and TensorRT, respectively.

From a technical perspective, GNU Radio must make some assumptions about scheduling and memory management that may limit the application's performance. As a result, some DSP developers (including Deepwave) find that, for certain applications, it makes sense to work with the software libraries that support GPUs natively (as shown in Figure 3). Most of these libraries

can be leveraged using Python with acceptable performance. For those wanting every last bit of performance, C++ interfaces are also provided.

In addition to the libraries provided by AirStack, NVIDIA has recently released an open source DSP toolbox, called cuSignal, as part of the RAPIDS accelerated data science project. cuSignal, by GPU accelerating the popular SciPy Signal library, demonstrates the capability for Python programmers to easily write GPU accelerated signal processing applications, making it even easier for DSP engineers to leverage the GPU. Deepwave is currently evaluating cuSignal for inclusion in future releases of AirStack and comparing it against our traditional workflow of using CUDA, cuFFT, and other software libraries directly.

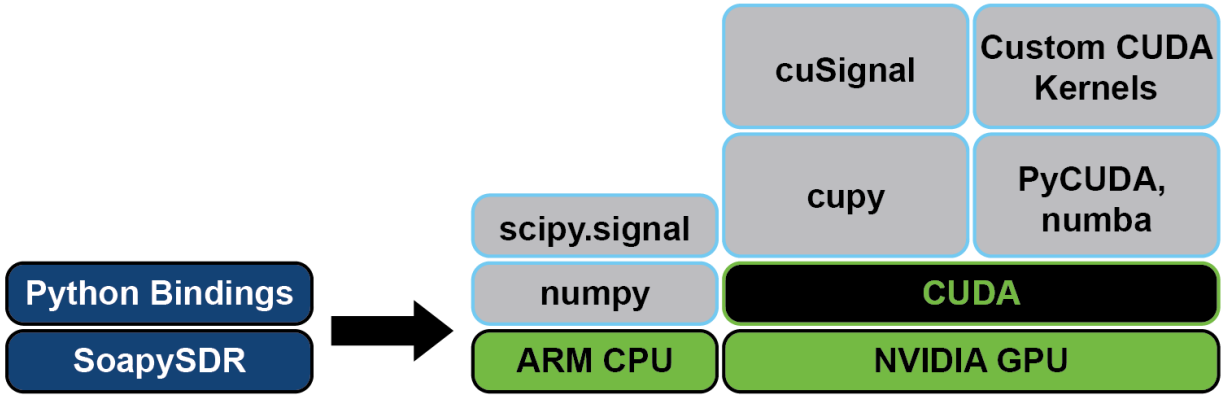

Figure 3: Python Software Suite for DSP on the AIR-T

3.2 Deep Learning Workflow with the AIR-T

The Artificial Intelligence (AI) training process for RF signals has similarities to more traditional AI application spaces, specifically both image processing and speech recognition. One major difference, however, is that RF signals are typically stored as a series of complex numbers (i.e., data with both a real and imaginary component) since this is a particularly convenient form for signal processing. While common in DSP, this data type is not natively supported by any deep learning framework. One method to work around this limitation is to perform a pre-processing step to transform the complex signal into a real-valued representation such as a spectrogram. Alternatively, the complex valued data stream may be treated as two real values. For this to work properly, the relationship between the real and complex number needs to be interpreted and learned during the training process. Either of these methods can be viable depending on the specific deep learning application.

Once a viable DNN model has been created (which can be done using any industry-standard machine learning framework), it may be deployed for inference using TensorRT. TensorRT is NVIDIA's DNN optimization and inference toolkit. The toolkit will optimize the model and produce a platform-specific frozen binary representation of the model. This optimization runs directly on the deployment hardware, e.g., the AIR-T. This full process is outlined in Figure 4.

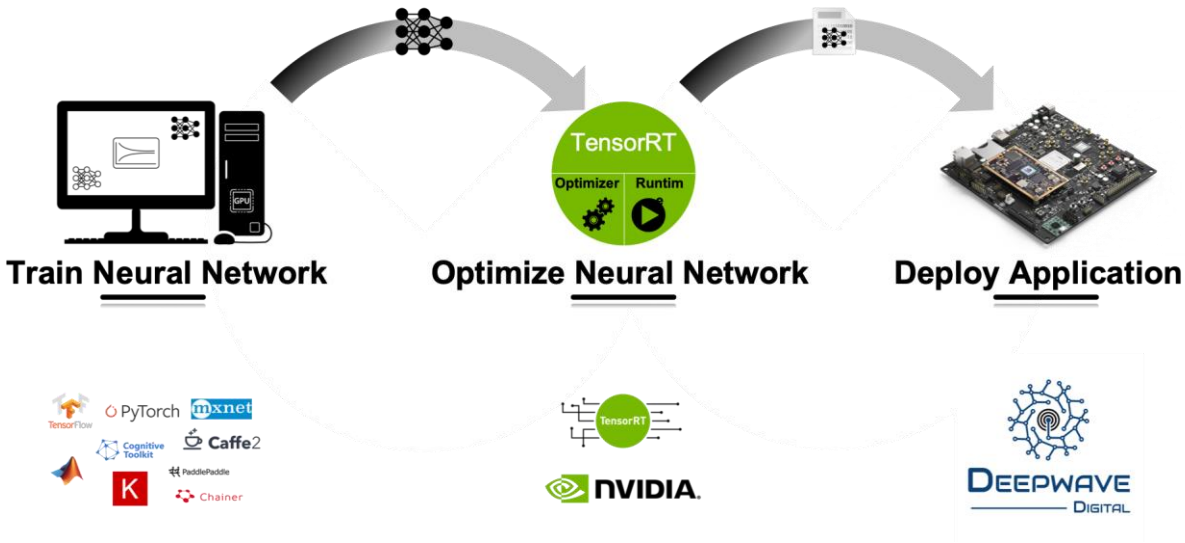

Figure 4: DNN deployment workflow.

4 CREATING AN ESC SENSOR

Deepwave worked with Key Bridge Wireless, a CBRS spectrum administrator, to develop a set of training data consisting of tens of thousands of radar signals spanning the entire CBRS parameter space, including various radar system designs, fluctuating power levels, and real-world over the air effects. The Deepwave CBRS signal classifier was trained on these data resulting in an enterprise-level signal classification solution that combines traditional DSP techniques (which are GPU accelerated) with advanced deep learning methods.

Referring to Figure 5, the AIR-T's receiver is tuned to the 3.5 GHz band and the signal stream is digitized in the receiver. This occurs using SoapyAIRT, an AIR-T specific plug-in for the open-source SoapySDR framework. The raw, complex-valued, digital signal is then transferred to the Jetson TX2, where 100% of the associated DSP processing occurs on the GPU. A key enabler here is the fact that the buffer used to read signal data from the RF frontend is already GPU addressable. Said differently, once the data arrives, no additional copy is required to get the data onto the GPU for further processing.

The first processing step takes the signal and separates it into a set of frequency channels so that they can be managed individually as part of the overall CBRS network design. This process, called channelization, is implemented on the Jetson's GPU as a series of CUDA kernels and creates the set of signals that will be fed to the classification algorithm.

Following channelization, a GPU-based detection algorithm (a CUDA kernel) is used to remove time periods where no signals exist that are strong enough to require protection, based on the CBRS specification. The remaining data is transferred (via a zero copy interface) to the DNN classification algorithm to determine if the detected signal is one of the protected transmissions from a priority user. If the classifier determines that the channel must be protected, then the AIR-T communicates this information back to the CBRS network management infrastructure.

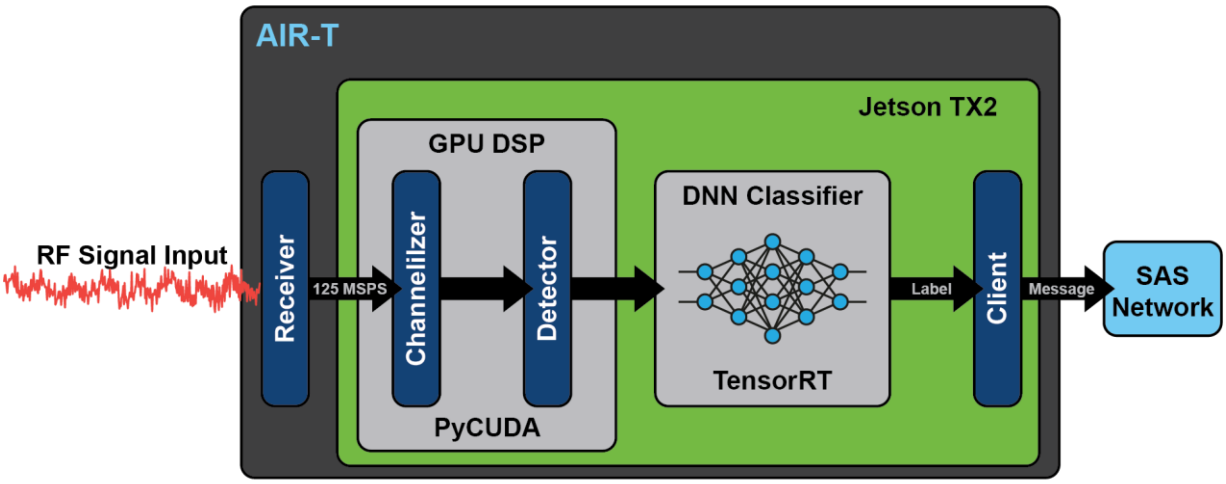

Figure 5: CBRS Signal Classifier

The Deepwave CBRS ESC network sensor provides a prime example of how enterprise-level solutions may be created, tested, and deployed on the Jetson product line and the AIR-T. This use case shows how traditional DSP can be combined with advanced deep learning algorithms to implement critical technology for next-generation telecommunications as well as many other industries that depend on signal processing. What's more, all the necessary software runs on a single Jetson TX2 and, by leveraging the onboard GPU, the resulting throughput is more than sufficient to process the required RF bandwidth.

Deployment for the Key Bridge Wireless ESC network, powered by the Deepwave Digital AIR-T and DNN, will be rolled out in 2020 and begin to offer service to enterprise customers. The network will be deployed along the coastline of the continental United States, Alaska, Puerto Rico, Guam, and Hawaii.

Please contact Deepwave Digital for inquiries and questions here.

Learn more about NVIDIA's developer resources for telecommunications here.

5 About The Authors

John Ferguson, CEO

John is CEO of Deepwave Digital, a startup enabling seamless integration of deep learning into edge compute wireless technology. He has significant experience developing algorithms for geolocation, deep learning, signal processing, radar, and communication systems. John holds dual B.S. degrees in Math and Physics from Appalachian State University and an M.S. and Ph.D. in Materials Science from Cornell University.

Peter Witkowski, Lead Software Engineer

Peter is an expert at developing drivers and software to interface with software defined radios. He has significant experience developing robust software for state-of-the-art wideband digital signal processing algorithms on GPU and CPU architectures. He received his B.S. in Computer Engineering from the University of Illinois and went on to get his M.S. in Systems Engineering from the Naval Postgraduate School.

William Kirschner, Lead FPGA Design Engineer

William Kirschner has significant experience developing communications signal processing algorithms for wireless, terrestrial, satellite, and cable systems. Specific areas of expertise include board, RF, ASIC, and FPGA design. He received dual B.S. degrees in Biomedical and Electrical Engineering from Syracuse University, an M.S.E.E. in Communications Theory from George Washington University, and an M.S.E.E. in signal processing from Villanova University.

Daniel Bryant, Lead Engineer

Daniel Bryant is an embedded software engineering specialist at Deepwave Digital with broad experience in integrating GPU computing into real-time sensing systems. At Deepwave, he is currently working on hardware platforms for machine learning with broadband radio signals. He is a graduate of Purdue University, receiving B.S. degrees in both Mathematics and Statistics.

6 REFERENCES

[1] "Deep Learning Classification of 3.5-GHz Band Spectrograms With Applications to Spectrum Sensing - IEEE Journals & Magazine." [Online]. Available: https://ieeexplore.ieee.org/document/8642956. [Accessed: 12-Dec-2019].

[2] "Key Bridge Wireless concludes CBRS ESC testing," *PRWeb*, 25-Nov-2019. [Online]. Available: https://www.prweb.com/releases/key_bridge_wireless_concludes_cbrs_esc_testing/prweb16743275.htm. [Accessed: 12-Dec-2019].