White Paper

Amplifying 5G vRAN Performance with Artificial Intelligence & Deep Learning

DeepSig's OmniPHY® 5G AI software combines Deep Learning with Intel® FlexRAN™ Reference Architecture and oneAPI Deep Neural Network Library to improve wireless performance and resource utilization of Distributed Unit (DU) Upper PHY software in 5G Open vRAN systems.

Authors

Dr. Tim O'Shea, CTO
DeepSig

Tong Zhang
Intel

Vitaliy Zakharchenko
Intel

Table of Contents

Introduction

5G Open virtualized Radio Access Network (Open vRAN) architecture enables operators to deploy best-in-class products from multiple suppliers. Today, Mobile Network Operators (MNOs) and private operators can deploy Intel® FlexRAN™ Reference Architecture to enjoy the benefits of Open vRAN, while using cutting-edge deep learning technology in the baseband to enhance performance and efficiency.

DeepSig's expertise in artificial intelligence (AI) and machine learning (ML) applied to wireless signal processing has enabled development of embedded software that replaces multiple 5G NR signal processing algorithms with a precisely designed Deep Neural Network (DNN). This approach with DNN potentially requires less computation while significantly improving network capacity and resilience to interference by learning the real-world characteristics of the local wireless environment where the Radio Unit (RU) operates. These improvements reduce both capital expenditure and operating expenses, which increases Open vRAN's value to MNOs. DeepSig and Intel collaborated to bring this transformational AI software to market as part of the Intel FlexRAN software suite.

Implementing the 5G NR radio access network is exponentially more complex than previous generations, particularly when seeking peak multi-user capacity gains afforded by Massive MIMO (mMIMO) and when applying many-element, multi-user processing techniques introduced in 5G. AI/ML advances deliver greater potential in 5G infrastructure by providing better service and capacity and by reducing computational load. This paper explains how DeepSig's initial 5G software release applies AI/ML uniquely within the upper-L1 baseband processing in the ORAN Distributed Unit (O-DU).

AI and Machine learning techniques are rapidly growing into core baseband signal processing, including the L1. Beyond the standards-transparent techniques described here, they will offer significant performance benefits in 5G advanced and 6G, increasingly relying on DNN processing to optimize RAN performance across the stack. While this work demonstrates immediate benefits within the Intel FlexRAN Reference Architecture 5G L1 stack, continued benefits will be realized in future versions, as well as through inclusion of deep learning and neural networks into Intel FlexRAN software as DNN enhance more traditional signal processing functions. Intel® Deep Learning Boost (Intel® DL Boost) processor extensions align perfectly to Deep Learning RAN (DL-RAN), providing low latency on-chip processor extensions to accelerate and optimize for DL inference. Additionally, the Advanced Matrix eXtensions (AMX) in the next-gen Intel® Xeon® processor – previously codenamed Sapphire Rapids – will continue to to increase neural network processing efficiency on Intel silicon, leading to an efficient convergence of RAN DSP and other signal processing applications all benefitting from enhanced DL inference extensions at scale.

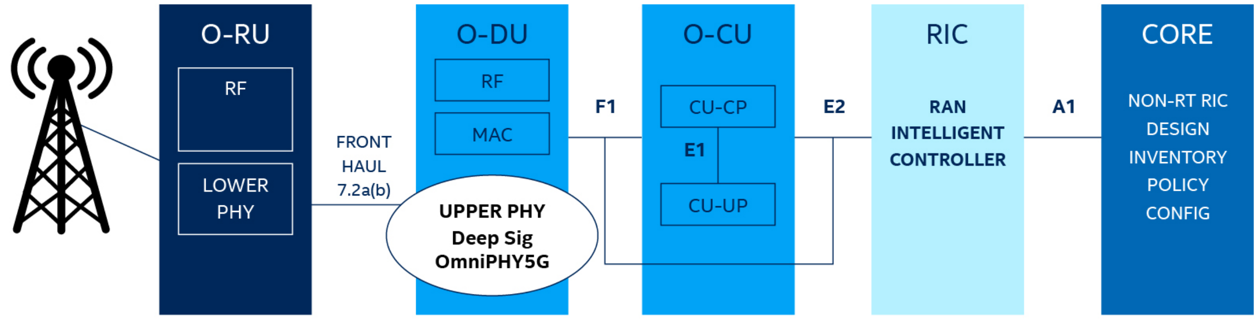


Figure 1: 5G Open vRAN Architecture Components DeepSig's Deep Neural Network is a software patch into Upper-L1 processing of the DU.

Background and Impact

5G Open vRAN leverages virtualized, general purpose compute resources from both Cloud and Edge-Cloud resource elements.

This design greatly reduces cost, simplifies resource management and allows for sharing resources between RAN and application execution. In the 5G Open vRAN architecture, the RAN is decomposed (see Figure 1) into a number of virtualized components, including the Radio Unit (RU), which contains the analog to digital conversion, amplifiers, antennas and low-L1 implementation. The Distributed Unit (DU) implements the upper-L1 processing including channel estimation, antenna processing and MAC/RLC processing. The Central Unit (CU) performs upper layer RLC and protocol processing for numerous sectors and the Core, which manages central routing, authentication, control, etc.

Intel® FlexRAN™ Reference Architecture

With the Intel FlexRAN Reference Architecture, Intel is offering a blueprint to quicken development of vRAN and Open RAN solutions, helping equipment manufacturers and operators reduce time, effort and cost. Intel FlexRAN Reference Architecture enables fast development of software-based LTE and 5G NR Base Stations that can be instantiated on any node of the wireless network from edge to core. The block diagram in Figure 2 shows the Intel FlexRAN software layer 1 (L1) PHY application, which takes radio signals from the RF front-end and performs real-time signal and physical layer processing on servers built with Intel® Xeon® Scalable processors. The architecture takes advantage of the Intel® Advanced Vector Extensions 512 (Intel® AVX-512) instruction set for efficient implementation of L1 signal processing tasks. The Intel FlexRAN Reference Architecture is optimized for NFVI cloud-native deployments by using DPDK based networking and the hardware capabilities of the Intel® Ethernet Network Adapters 700/800 series. The Intel FlexRAN Reference Architecture performs the entire 4G and/or 5G layer 3, 2, and 1 processing and utilizes Intel dedicated hardware accelerators for FEC (Intel® FPGA Programmable Acceleration Card N3000 or Intel® vRAN Dedicated Accelerator ACC100 Adapter) as well as cryptographic accelerators (Intel® QuickAssist Technology). This in turn provides more processing power available to increase cell capacity and edge-based services and applications.

This approach is fundamentally about building 5G mobile networks using a fully programmable software-defined solution based on open interfaces that run on commercial off-the-shelf hardware (COTS) with open interfaces.

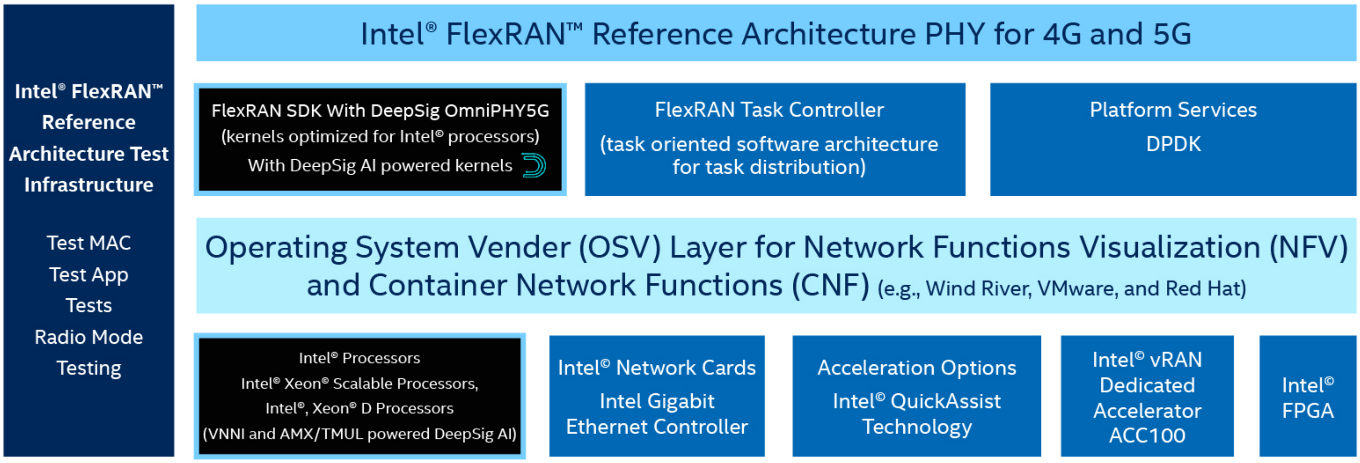

Intel® FlexRAN™ Reference Architecture Test Infrastructure
Test MAC
Test App
Tests
Radio Mode
Testing
Intel® FlexRAN™ Reference Architecture PHY for 4G and 5G
FlexRAN SDK With DeepSig OmniPHY5G
(kernels optimized for Intel® processors)
With DeepSig AI powered kernels
FlexRAN Task Controller
(task oriented software architecture for task distribution)
Platform Services
DPDK
Operating System Vender (OSV) Layer for Network Functions Visualization (NFV) and Container Network Functions (CNF) (e.g., Wind River, VMware, and Red Hat)
Intel® Processors
Intel® Xeon® Scalable Processors,
Intel®, Xeon® D Processors
(VNNI and AMX/TMUL powered DeepSig AI)
Intel® Network Cards
Intel Gigabit Ethernet Controller
Acceleration Options
Intel® QuickAssist Technology
Intel® vRAN Dedicated Accelerator ACC100
Intel® FPGA

Figure 2: Intel FlexRAN Reference Architecture Showing DeepSig Additions to the Intel FlexRAN software and Drop-In Usage in Reference PHY.

Historically computational complexity of signal processing of Physical layer required designing costly custom silicon SoCs where PHY layer algorithms instantiated are tightly coupled between hardware capability of silicon and corresponding software components. This has changed with the introduction of Intel FlexRAN Reference Architecture, where general purpose CPUs are used to deliver PHY layer signal processing tasks in an efficient manner.

This architecture enables:

1. disaggregation of hardware and software, benefiting speed of innovation cycles for both.
2. networks that are more open, and thus a broader ecosystem of vendor options for MNOs
3. the movement of previously silicon-specific design questions and capability to a flexible and agile software layer, with cheaper research and development cost
4. reduced time to market for Telecom Equipment Manufacturers (TEMs), replacing years of silicon design by allowing faster development of software upgrades

With this type of disaggregation, hardware used for instantiation of Base station nodes in MNO 5G network becomes the same as the rest of 5G network nodes between the Data Center, Core Network and Edge of the Network. Using a PHY layer in software makes the O-DU node fully software defined, which unlocks homogeneity of the entire 5G network to be defined in software, programmable from end-to-end, as well as in a unified control plane that controls it from top to bottom for all types of nodes.

In addition to these benefits, orchestration of all network functions within the 5G network becomes possible using the same Cloud Native approaches. MNOs don't need to deal with RAN nodes in a separate custom fashion anymore. This in turn helps to reduce overall management and operations costs and benefits economies of scale for deployment of RAN networks on the same unified hardware platform.

Based on the deep learning revolution started from seminal work on ImageNet for computer vision in 2012[5], proliferation of machine learning algorithms into different domains and type of workloads is wide, robust and on an accelerating trend. Wireless commutation and PHY specifically are no exception here.

A broad list of academic institutions, standards bodies such as 3GPP and commercial companies apply data-driven ML approaches to 5G and PHY specific tasks. As a result, this becomes fruitful ground for significant innovations in wireless network operations, both today and in the future. Effective instantiation of ML inference within the PHY pipeline of the O-DU node of 5G networks becomes an important topic of consideration. On a technical level, this approach combines the complexity of classical signal processing algorithms with the complexity of machine learning techniques and data driven methods. Given the innovation rate of ML models alone, as well as specifics and constrains of PHY layer, an inference custom silicon SoC platform can become outdated pretty much on the day silicon design starts. However, COTS platforms are updated on a predictable cadence. Unlike purpose-built processors, they are designed for use cases requirements across many domains where ML inferencing is top of mind, such as Public Clouds, IOT and general software-defined networking. Selecting the right platform for the next generation of RAN becomes a very important and non-trivial task. It is easy to see that the cadence of innovation will accelerate, and a platform such as the CPU provides a reliable tool to address these types of challenges. In addition, continued advancements in instruction sets specific to AI/ML, as well as the growing ecosystem of ML-specific and wireless-specific tools, libraries and frameworks will be available to users as part of open source and Intel FlexRAN technology future releases. The unparalleled ecosystems of both wireless and machine learning give TEMs and MNOs the ability to focus on innovative work instead of dealing with the limitations of hardware capabilities and long design cycles of custom silicon. Other computation platforms that might be beneficial for ML acceleration itself are not a very good fit for RAN. Very often, these platforms simply are not deployable in RAN networks due to limits of size, power consumption and price. One of the greatest benefits of the CPUs is the preservation of homogeneity of the 5G network from Core to Edge to RAN in terms of hardware platform and full readiness for ML-specific algorithms.

DeepSig 5G AI

DeepSig's 5G AI embedded software provides a set of enhancements to the Intel FlexRAN software for DU, which fits into the existing Intel FlexRAN software development kit (SDK) and software architecture as shown in Figures 2 and 3. This provides drop-in replacements to the PUSCH channel estimation SDK routines (for standard MIMO), and to SRS channel estimation and pre-coding routines (for mMIMO). DeepSig's 5G AI software components can be readily leveraged by existing Intel FlexRAN software for DU vendors without the need for any additional changes in their hardware or software stack. Additionally, cloud services provide online learning and adaptation of these routines to continually enhance performance over time. Real-world data may be deployed alongside the runtime components in future versions co-located on the DU, in nearby RIC xAPPs or on other mobile edge compute platforms.

The most critical performance considerations in 5G vRAN systems are power consumption, computational cost of processing radio units, spectral efficiency and realized throughput for mobile users. These crucial performance areas are dramatically enhanced through machine learning-based processing approaches within the DU, and specifically by the unique way that OmniPHY-5G changes Upper-L1 processing in the DU using a data-driven DNN. This results in two key benefits: processing efficiency and signal-to-interference and noise (SINR) improvements.

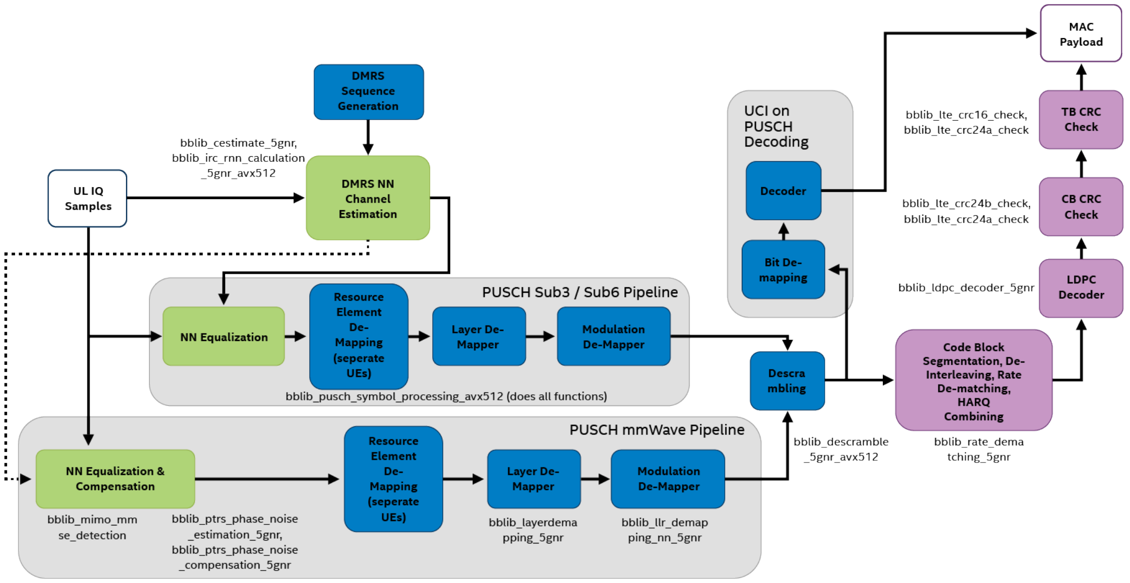

DMRS Sequence Generation
bblib_cestimate_5gnr, bblib_irc_rnn_calculation _5gnr_avx512
UL IQ Samples
DMRS NN Channel Estimation
PUSCH Sub3 / Sub6 Pipeline
NN Equalization
Resource Element De-Mapping (seperate UEs)
Layer De-Mapper
Modulation De-Mapper
bblib_pusch_symbol_processing_avx512 (does all functions)
PUSCH mmWave Pipeline
NN Equalization & Compensation
Resource Element De-Mapping (seperate UEs)
Layer De-Mapper
Modulation De-Mapper
bblib_mimo_mmse_detection
bblib_ptrs_phase_noise _estimation_5gnr, bblib_ptrs_phase_noise _compensation_5gnr
bblib_layerdemapping_5gnr
bblib_llr_demapping_nn_5gnr
Descrambling
bblib_descramble _5gnr_avx512
UCI on PUSCH Decoding
Decoder
Bit De-mapping
Code Block Segmentation, De-Interleaving, Rate De-matching, HARQ Combining
bblib_rate_dematching_5gnr
LDPC Decoder
bblib_ldpc_decoder_5gnr
CB CRC Check
bblib_lte_crc24b_check, bblib_lte_crc24a_check
TB CRC Check
bblib_lte_crc16_check, bblib_lte_crc24a_check
MAC Payload

Figure 3: Intel® FlexRAN™ Reference Architecture PUSCH Processing Flow Graph Augmented with DeepSig NN Estimation, Equalization, and Compensation.

By using an ML approach to L1 processing, less compute is required to process the uplink (PUSCH) signals in standard MIMO configurations, reducing the cost of operation and increasing the number of sectors per server.

OmniPHY-5G can attain significant improvement in computational kernel latency reduction in PUSCH processing on Intel Xeon Scalable processors in standard MIMO modes as evaluated on 3GPP TDL channel models for 38.104 testing. The ML-driven L1 processing improves SINR performance, a benefit which enables increased throughput and coverage, while optimizing the value and utilization of costly spectrum licenses and band allocations.

Because OmniPHY-5G can attain SINR improvements, MNO can benefit from bandwidth increases, user traffic latency reduction and the reduction of necessary interference margin for cell planning.

Impact

DeepSig's 5G AI initial demonstration version provides overall computational efficiency improvement for a typical deployment scenario, along with throughput enhancement due to improved SINR. Leveraging existing DU hardware and DeepSig's 5G AI provides significant notional cost per-bit reduction when considering both factors for 4 antenna systems. These critical AI software enhancements make Open vRAN significantly more performant and competitive to traditionally designed and optimized vRAN solutions.

Approximated maximum throughput rates can be derived from the Shannon-Hartley theorem[4], which relates achievable data-rate with SINR for a specific channel width. By comparing relative throughputs for different SINR levels, the throughput percent increase impact for a range of baseline SINR levels is shown in Figure 6, with all other factors kept constant (i.e., channel bandwidth or resource block allocation size). Specifically, in cell-edge and low SINR cases, this can significantly help with throughput and wireless quality of service and allows for the use of higher modulation and coding rates (MCS) in many access situations to improve spectral efficiency. Figure 4 shows an example of the expected throughput impact for the range of measured SINR gains over a range of operating SINRs that may be seen in common UE deployment scenarios.

Intel Xeon processor architecture is continuing to improve DL and signal processing performance as architecture, execution, memory and bandwidth capacities grow with each generation. While most results in this paper are conducted on 2nd Gen Intel Xeon Scalable Processors SKU 6248 @ 2.5 GHz, Figure 5 shows the impact of an equivalent 3rd Gen Intel Xeon Scalable Processor SKU 6338N also with VNNI (DL Boost) extensions at identical clock speeds. The graph shows 5-25% improvements in each case when moving to 3rd Gen, and significant further gains are expected on next-generation Intel Xeon processors with both VNNI and TMUL/AMX extensions in the future[1].

Intel® Xeon® Scalable Processors

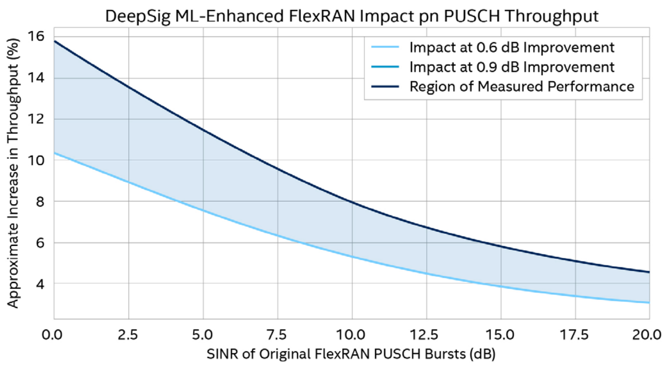
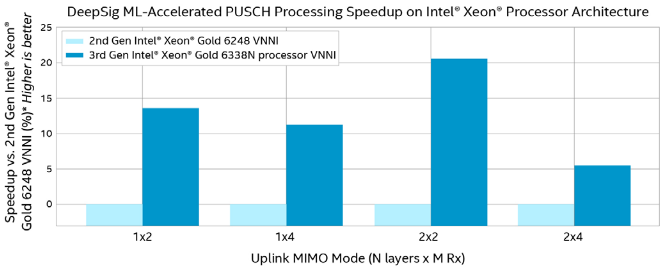


Figure 4: Example of Notional Impact of OmniPHY on PUSCH Throughput.



* Zero is baseline, not absolute performance. For exact specifications, see Footnote 1.

Figure 5: Neural Network Acceleration by Intel® Xeon® Processor Architecture.

Scaling to Massive MIMO

mMIMO is rapidly gaining adoption within 5G networks and deployments due to its unprecedented spectral efficiency, multi-user capacity and improved coverage and throughput. However, improved implementations are needed to reduce cost, improve performance and accelerate deployments. While DeepSig's AI driven L1 optimization techniques described herein are initially provided for standard MIMO processing, the channel estimation, equalization and combining algorithms optimized here represent an even larger portion of the computational time and cost within mMIMO. An enhanced 5G AI version for mMIMO L1 processing solution is under development and tests in DeepSig's 5G AI Lab to provide even greater benefit in these key processing sections of mMIMO.

DeepSig expects Intel FlexRAN software's computational reduction for a fully-loaded gNB in 4TRx mode SINR improvement obtained with standard MIMO to provide a substantial computational reduction in mMIMO processing (considering only the SRS and PUSCH acceleration cases), while providing additional SINR margin improvements, especially when including online learning in future releases. These improvements are only the first step in optimizing mMIMO with AI/ML. While 2TRx and 4TRx 5G-NR MIMO systems today use purely accelerated DMRS/PUSCH processing, this can change in 32TRx and 64TRx mMIMO modes. As a result, implementation of the mMIMO receiver can be notoriously complex and technically difficult to make efficient. Speedup from usage of ML in mMIMO can be expected from next areas: (1) how the SRS processing will be accelerated by ML, (2) how both SRS and PUSCH processing will both be accelerated together in a similar fashion, and (3) an approach where all three SRS, PUSCH and beamforming weight calculation (BF) are accelerated using a similar ML approach to provide even more significant compute reductions for Intel FlexRAN Reference Architecture's Upper L1 without any addition or modification to hardware.

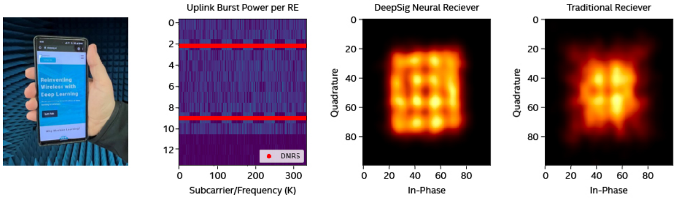


Figure 6: DeepSig 5G-AI Lab, Hardening ML Software Over the Air.

Testing and Validation

In the first quarter of 2021, DeepSig opened its 5G Wireless AI Lab using commercial products to construct an end-to-end 5G SA network based on Open vRAN architecture. With a mid-band FCC experimental license, DeepSig conducted 5G NR over-the-air (OTA) tests and model validations. Ongoing tests are demonstrating the ML efficacy and how it performs even better OTA than in DeepSig's 3GPP measurements. Figure 6 illustrates a commercial 5G UE attaching to the DeepSig 5G system SA network and the internet. These measurements are taken along with signal quality performance numbers obtained using traditional estimation and equalization approaches (MMSE) alongside the neural network (NNEQ) approach to estimation and equalization OTA while running standard commercial UE apps and data services.

ML-Based Channel Processing

Key challenges

ML presents a key opportunity within the 5G and beyond-5G vRAN to perform enhanced channel estimation, equalization and combining of maximal ratio. ML enables telcos to better preserve information versus today's widely used implementations and does so at low complexity in order to minimize capital and operating costs. A high-level architecture of the neural network-based approach is provided in Figure 7 to illustrate how and where this solution is realized.

Model Architecture

DeepSig's model architecture has been designed with AutoML to take advantage of efficient Intel AVX-512 extensions and optimized neural network layer primitives available within oneDNN. The primitives include convolutional layers, rectified linear units and numerous state-of-the-art network architecture and training techniques, which enable an inference architecture that is flexible to different PUSCH burst configurations. These are compact and have low complexity and latency to execute, and are accurate to provide excellent SINR, BER and FER statistics, resulting in improved user quality of experience (QoE). In contrast to many large computer vision networks, this solution can process input PUSCH resource block allocations and DMRS tones and attain these goals within a very compact architecture. The inputs can be executed in as little as 19 microseconds to recover equalized symbols and channel state, accelerating and improving Open RAN performance on Intel Xeon processors.

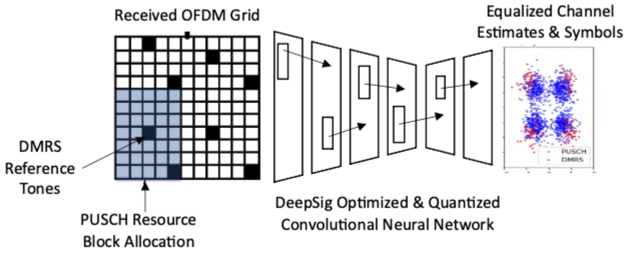

Received OFDM Grid
Equalized Channel Estimates & Symbols
DMRS Reference Tones
PUSCH Resource Block Allocation
DeepSig Optimized & Quantized Convolutional Neural Network
PUSCH
DMRS

Figure 7: High Level ML Estimation and Equalization Pipeline.

Intel® Xeon® Scalable Processors

All 2nd Gen Intel Xeon Scalable processors and later feature Intel AVX-512 and Intel DL Boost, which can help accelerate machine-learning training and inference. 3rd Gen Intel Xeon Scalable processors feature enhanced Intel DL Boost with the industry's first x86 support of Brain Floating Point 16-bit (bfloat16 supported on 4 socket parts) and Vector Neural Network Instructions (VNNI), which provide enhanced AI inference and training performance. Depending on the AI workload, for example, these technologies in the 3rd Gen Intel Xeon Scalable processor can deliver up to 1.93 percent more AI training performance[2] and 1.87X more AI inference performance[3] compared to the previous generation.

ML Performance Generalization and Deployment Considerations

The performance measurement comparisons shown for differing Standard-MIMO antenna configurations in Figure 5 are conducted using industry standard 3GPP TDL channel models for machine learning model training and test, which are equivalent to those used for radio frequency conformance tests (RCTs). These training and test data sets are large and distinct (i.e., not re-used), but they are drawn from the same statistical distribution of the channel model (i.e., fixed parameters of each standard channel fading profile such as TDL-A-30-10, TDL-B-100-400, TDL-C-300-100) along with randomized PUSCH frame data and DMRS references to avoid any unfair model overfitting.

One key consideration in all ML systems is that of "generalization," i.e., how well the training data and training process enable the resulting inference ML model to perform on real data that will be seen during deployment at inference time. In some cases, performance can be degraded in a machine learning model if the generalization is poor. This is a major and often overlooked problem in early proof of concept works which are not trained, tested and hardened with fair and realistic assumptions. It is also a key focus of DeepSig's efforts in hardening this ML approach for production.

Ensuring generalization and performance in production is a critically important consideration when comparing ML-based approaches, such as this one, to closed form expressions such as the MMSE equalizer, which does not rely on specific training data sets. Instead, they encapsulate their own set of assumptions at design time. Therefore, closed-form approaches such as MMSE are not expected to encounter this form of generalization issue as long as the design-time assumptions hold. The closed-form approaches are unable to adapt and leverage additional statistical information after design-time to improve performance from data in the way that ML approaches can.

Results from OTA measurement in Figure 6 are a key point in this work to demonstrate the ability of the ML approach. The model used here is similarly trained on a standard 3GPP TDL fading model in simulation, but then evaluated using OTA data collected from DeepSig's 5G SA n78 Intel FlexRAN software-based OpenRAN system. The system uses PUSCH emissions from commercial UE containing previously unseen data and channel responses (and in fact, also an unseen slot configuration, resource block allocation size and DMRS configuration). In this case, significant performance advantages are retained with the approach in both inference time and attained SINR. This key result shows that ML models can be used and generalize well in some cases.

Generalization and real-world tests will continue to be a critical consideration when comparing adoption of ML-based vs. non-ML-based approaches in L1 signal processing as ML methods are increasingly adopted to take advantage of their attractive performance benefits. Extensive and rigorous testing and realistic operating assumptions will continue to be critical for vetting the performance of ML-based approaches across a wide range of conditions. As with many applications of ML, mature validation, test, generalization capabilities and tracking and improvements of failure modes and conditions over time will be important to continue to harden the approach and ensure its performance in production. To address hardening of DeepSig's approach against all real-world operating conditions, a number of techniques are employed, including:

- Fail safes
- Online model training and management of channel data and models across various environments in the network
- Continued improvements to the ML models and training processes
- Data handling processes to continue improvements to generalization robustness and to performance over time

The techniques underscore focus on evaluation and measurement with 5G hardware in the loop, rather than pure reliance on statistical models. 3GPP TDL channel models (or trivial Gaussian or Rayleigh channel models used by some other academic works) do not represent all possible channel responses and can be seen in real world operating deployments. DeepSig will continue to scale and accelerate these aspects of OmniPHY 5G and ML Model enhancement and validation rapidly over the coming months. Continued field testing and field hardening of the model, training and deployment components will continue to mature as these efforts expand to ensure model generalization and model performance in a wider range of real-world operating conditions, ensuring performance and robustness of the approach.

ML Software Platform

Training and Validation Data Generation

Data is the core of any machine learning approach. DeepSig engaged several sources to provide effective, high-speed training and validation against known vetted receiver compliance and 3GPP test cases and validation on OTA test data in order to cover all three critical areas. The DeepSig 5G SIM delivered high-speed data generation, while Intel FlexRAN software's existing Radio Conformance Tests (RCT) cases supported validation. Lastly, Deep Sig's 5G SA lab system provided initial OTA validation and test.

Training

Training couples high-speed generation of NR frame data and channel simulation plus effects with state-of-the-art neural network design. The DeepSig EQ Training process uses PyTorch to train a compact custom convolutional neural network (CNN) architecture to provide both channel estimation and equalization functions within the L1 on uplink DMRS reference tones within the PUSCH. Then the DeepSig AutoML Optimizer assists with architecture and parameter selection and an INT8 Quantization tool built on top of PyTorch to obtain a highly performant neural network.

Inference

High performance inference utilizes the DeepSig Realtime Inference library, which leverages Intel® oneAPI Deep Neural Network Library (oneDNN) inference kernels and provides low-latency accurate full network inference. It is then linked directly into the Intel FlexRAN software, where it can be used by the Intel FlexRAN software L1 App, TestApp and other test cases.

Inference Hardware Considerations

The way the ML model is instantiated in the O-DU system has significant impact on system-level performance indicators, such as inference time and O-DU power consumption. It can also provide additional requirements for the hardware platforms that deploy the Radio Access Network. In addition, ML training and optimization steps can be quite different depending on what target hardware block performs ML inference in the system (i.e., CPU, GPU, FPGA, dedicated ML accelerator). This requires additional engineering effort to fine tune the model for given hardware blocks and can increase overall complexity, time to market and cost. Overall benefits obtained from deploying the ML model for channel estimation and equalization should be evaluated against those important factors. Total cost of ownership (TCO) of a deployed O-DU solution may be impacted based on the approach taken.

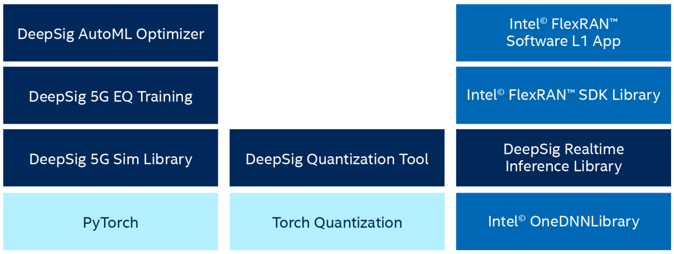

DeepSig AutoML Optimizer
DeepSig 5G EQ Training
DeepSig 5G Sim Library
PyTorch
DeepSig Quantization Tool
Torch Quantization
Intel© FlexRAN™ Software L1 App
Intel© FlexRAN™ SDK Library
DeepSig Realtime Inference Library
Intel© OneDNNLibrary

Figure 8: Training, Quantization, and Inference software stack.

The ideal approach for Independent Software Vendors (ISVs), as well as MNOs, would be to get the benefits of AI/ML without any additional downside to key parameters of the O-DU as it is deployed in the field. This can be achieved today with Intel Xeon Scalable processors, where the same software engineering techniques and tools used for field-proven Intel FlexRAN software-based deployments can be used to instantiate the very best ML methods in a high-performance and cost-effective way. Adding extra hardware components specifically for ML model inference introduces more inefficiencies on several levels: extra complexity of project development phase; extra complexity of deployment configurations; degradation of performance per watt and performance per unit of space; and power consumption increase. As a result, TCO is significantly impacted. When it comes to ML, VNNI instruction sets are already available and deployed in most of Open vRAN deployments and have proven very effective for ML inference in this work.

Solution Validation

Validation is important on many levels for the use of AI/ML within the L1 and the receive chain, which must be resilient and performant during all operating conditions. First, validation is run on existing Intel FlexRAN software 5G NR RCT tests using existing TestAPP and Test MAC infrastructure to ensure compliance with all existing tests. Second, tests use a high-speed simulator, which produces millions of 3GPP-compliant frame configurations and millions of random 5G-NR TDL channel instantiations to provide complete channel performance validation across billions of possible operating modes. Finally, test and validation are completed on the ML-driven L1 on top of commercial Open vRAN hardware and software stacks to ensure proper operation over the air with hardware-in-the-loop. This arrangement allows validation and measurement of performance with a commercial UE (CUE), putting the software and algorithms to the ultimate test under local harsh, urban mid-band operating conditions.

Field-deployed live macro-cell testing and performance validation begins in 2022, and customer trials will follow. Also, integration is underway in DeepSig's 5G Wireless AI Lab, with additional Open vRAN hardware and software stack components to validate more end-to-end architectures and bring AI-native performance benefits to public and private Open vRAN mobile network operators.

Conclusion

This joint effort by DeepSig and Intel has demonstrated a unique and effective approach to upper PHY acceleration in the DU on Intel Xeon processors to enhance virtualized RAN performance through software upgrade alone. This approach will benefit both computational and signal quality, resulting in decreased cost-per-bit in some cases today, along with improved link-level margins for deployments. The test results are expected to improve on next-gen Intel Xeon architectures, which are especially optimized to enhance neural network inference performance and data movement.

The Intel Xeon Scalable processor family has many features, including Intel AVX-512 and Intel DL Boost. They also feature software tools, such as oneDNN library and optimized version of ML frameworks such as PyTorch. These tools provide a powerful and convenient environment for RAN software vendors to develop, train and deploy AI/ML-improved wireless solutions for Open RAN/vRAN networks.

DeepSig's OmniPHY AI-enhanced DU processing has shown how it can dramatically reduce the TCO and improve the performance and user experience of 5G vRAN deployments. It accomplishes this through the use of deep learning within the very low latency DU algorithms and by better exploiting data to improve baseband processing. While the focus has been on improving the DU in 5G vRAN with transparent and 3GPP-compliant processing, AI holds the promise of further improving the efficiency and performance of RU processing, fronthaul transport, low latency CU resource control and higher latency RIC-based resource and network control and allocation. Finally, AI has become increasingly recognized as the key enabler for 6G RAN, and DeepSig strongly believes the path to 6G begins with incrementally leveraging AI/ML within 5G vRAN. By continually leveraging more data and enhancing additional functions in real world systems, it is expected that incremental, robust and low-cost software upgrades will play an important role in the evolutionary path to 6G.

About

DeepSig Inc. is a venture-backed and product-centric technology company developing revolutionary wireless software solutions using unique, high-performance machine learning techniques to transform critical baseband processing tasks, wireless sensing and other key wireless applications.

Learn More

Intel® Xeon® Processors: https://www.intel.com/xeon

One-DNN: https://github.com/oneapi-src/oneDNN

DeepSig: https://www.deepsig.ai/

References

1. Configuration Details: Cascade Lake: 2nd Gen Intel® Xeon® Gold 6248 @ 2.50GHz processor on Intel Reference Platform (Wolf Pass) with 128 GB (12 slots / 16GB / 2667) total memory, ucode 0x500002c, HT on, Turbo on, with CentOS 7.8.2003, Linux* 3.10.0-1127.19.1.rt56.1116.el7.x86_64, Intel 800GB SSD OS Drive; Ice Lake: 3rd Gen Intel® Xeon® Gold 6338N @ 2.20GHz processor on Intel Reference Platform (Coyote Pass) with 128 GB (12 slots / 16GB / 2667) total memory, ucode 0xd000270, HT on, Turbo on, with CentOS 7.8.2003, Linux* 3.10.0-1127.19.1.rt56.1116.el7.x86_64, Intel 800GB SSD OS Drive; with Deepsig AI miniCNN train with Torch v1.7.0, inference running with oneAPI Deep Neural Network Library (oneDNN) v2.3 https://github.com/oneapi-src/oneDNN commit# 593e0de6267d2575f3e4c9e9818f0f11253d093a using Xbyak (5.993); JIT assembler for x86(IA32), x64(AMD64, x86-64) by C++ https://github.com/herumi/xbyak , test by Intel on 7/10/2021. https://www.intel.com/content/www/us/en/products/docs/processors/xeon/3rd-gen-xeon-scalable-processors-brief.html

2. Up to 1.93x higher AI training performance with 3rd Gen Intel® Xeon® Scalable processor supporting Intel® Deep Learning Boost (Intel® DL Boost) with BF16 vs. prior generation on ResNet50 throughput for image classification – New: 1-node, 4x 3rd Gen Intel® Xeon® Platinum 8380H processor (pre-production 28C, 250W) on Intel Reference Platform (Cooper City) with 384 GB (24 slots / 16GB / 3200) total memory, ucode 0x700001b, HT on, Turbo on, with Ubuntu* 20.04 LTS, Linux* 5.4.0-26,28,29-generic, Intel 800GB SSD OS Drive, ResNet-50 v 1.5 Throughput, https://github.com/Intel-tensorflow/tensorflow -b bf16/base, commit #828738642760358b388d8f615ded0c213f10c9 9a, Modelzoo: https://github.com/IntelAI/models/ -b v1.6.1, Imagenet dataset, oneAPI Deep Neural Network Library (oneDNN) 1.4, BF16, BS=512, test by Intel on 5/18/2020. Baseline: 1-node, 4xIntel® Xeon® Platinum 8280 processor on Intel Reference Platform (Lightning Ridge) with 768 GB (24 slots / 32GB / 2933) total memory, ucode 0x4002f00, HT on, Turbo on, with Ubuntu* 20.04 LTS, Linux* 5.4.0-26,28,29-generic, Intel 800GB SSD OS Drive, ResNet-50 v 1.5 Throughput, https://github.com/Intel-tensorflow/tensorflow -b bf16/base, commit #828738642760358b388d8f615ded0c213f10c99a, Modelzoo: https://github.com/IntelAI/models/ -b v1.6.1, Imagenet dataset, oneAPI Deep Neural Network Library (oneDNN) 1.4, FP32, BS=512, test by Intel on 5/18/2020. https://www.intel.com/content/www/us/en/products/docs/processors/xeon/3rd-gen-xeon-scalable-processors-brief.html

3. Up to 1.87x higher AI Inference performance with 3rd Gen Intel® Xeon® Scalable processor supporting Intel® Deep Learning Boost (Intel® DL Boost) with BF16 vs. prior generation using FP32 on ResNet50 throughput for image classification – New: 1-node, 4x 3rd Gen Intel® Xeon® Platinum 8380H processor (pre-production 28C, 250W) on Intel Reference Platform (Cooper City) with 384 GB (24 slots / 16GB / 3200) total memory, ucode 0x700001b, HT on, Turbo on, with Ubuntu* 20.04 LTS, Linux* 5.4.0-26,28,29-generic, Intel 800GB SSD OS Drive, ResNet-50 v 1.5 Throughput, https://github.com/Intel-tensorflow/tensorflow -b bf16/base, commit #828738642760358b388d8f615ded 0c213f10c99a, Modelzoo: https://github.com/IntelAI/models/ -b v1.6.1, Imagenet dataset, oneAPI Deep Neural Network Library (oneDNN) 1.4, BF16, BS=56, 4 instances, 28-cores/instance, test by Intel on 5/18/2020. Baseline: 1-node, 4x Intel® Xeon® Platinum 8280 processor on Intel Reference Platform (Lightning Ridge) with 768 GB (24 slots / 32 GB / 2933) total memory, ucode 0x4002f00, HT on, Turbo on, with Ubuntu* 20.04 LTS, Linux 5.4.0-26,28,29-generic, Intel 800GB SSD OS Drive, ResNet-50 v 1.5 Throughput, https://github.com/Intel-tensorflow/tensorflow -b bf16/base, commit #828738642760358b388d8f615ded0c213f10c99a, Modelzoo: https://github.com/IntelAI/models/ -b v1.6.1, Imagenet dataset, oneDNN 1.4, FP32, BS=56, 4 instances, 28-cores/instance, test by Intel on 5/18/2020. https://www.intel.com/content/www/us/en/products/docs/processors/xeon/3rd-gen-xeon-scalable-processors-brief.html

4. Shannon, Claude Elwood. "Communication in the presence of noise." Proceedings of the IRE 37.1 (1949): 10-21

5. Krizhevsky, A., Sutskever, I. and Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks NIPS 2012: Neural Information Processing Systems, Lake Tahoe, Nevada

Notices & Disclaimers

Performance varies by use, configuration and other factors. Learn more at www.intel.com/PerformanceIndex
Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or componen can be absolutely secure.
Your costs and results may vary.
Intel technologies may require enabled hardware, software or service activation.
Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.