Case Study

3rd Gen Intel® Xeon® Scalable processors
Speech Analysis and Synthesis

Custom Vocoder Optimization Solution for Enhanced Real-time Speech Synthesis

The new generation Intel® Xeon® Scalable processors empower Tencent Cloud's Xiaowei intelligent speech and video service access platform.

腾讯云小微

"Continuous improvements to throughput and real-time performance have made it possible for the Xiaowei platform to provide high-quality intelligent speech services for enterprise-grade applications. With the support of the advanced hardware and software technologies from Intel, the custom solutions based on the 3rd Gen Intel® Xeon® Scalable processors have unleashed the platform's speech synthesis performance."

Qiao Tian
Senior Researcher
Tencent Cloud

Intelligent speech applications are undergoing unprecedented breakthroughs and growth. The Chinese intelligent speech market is expected to reach CNY 19.48 billion by the end of 2021[1]. Tencent has been dedicated to artificial intelligence (AI) research and Internet innovations to empower intelligent speech hardware vendors. The company is currently working hard on the development of the Xiaowei intelligent speech and video service access platform. The platform, with Text to Speech (TTS) based on neural-based vocoder at its core, performs high-quality TTS conversion and delivery via end-to end acoustic models.

While classic vocoder models such as WaveNet can generate high-fidelity audio, the high complexity and huge computation required lengthen the synthesis of speeches, limiting their ability to satisfy the demand for real-time performance in real-world production scenarios. Continued access by a large number of devices also challenges the platform's throughput. Expanding server capacity is simply an imperfect solution, as it would cause deployment costs to skyrocket. For that reason, Tencent decided to adopt even more cutting-edge vocoder models to optimize the Xiaowei platform in-depth. In close collaboration with Intel, Tencent developed the Parallel WaveNet (pWaveNet) and WaveRNN custom vocoder model solutions to provide the platform with exceptional TTS performance while effectively reducing the total cost of ownership (TCO).

The solutions use the 3rd Gen Intel Xeon Scalable processors as its core computing engine. In addition to the increased cores and threads, which equip the platform with stronger computing capabilities, the next-generation processor is also integrated with BFloat (BF16) extensions and Intel® Advanced Vector Extensions 512 (Intel® AVX-512), which greatly reduces access to memory and supports hardware acceleration when working in conjunction with the Intel® oneAPI Deep Neural Network Library (oneDNN). The processor's larger cache also helps improve processing efficiency through higher cache hit rates. The custom solutions integrating the abovementioned advanced Intel technologies have enabled the Xiaowei platform to offer world-class speech synthesis performance to enterprises and device vendors. The feedback from the market has been extremely positive.

The custom vocoder model optimization solution delivered significant enhancements to Tencent's Xiaowei platform.

Faster Response - The custom pWaveNet vocoder model is advantageous in parallel computation through simplified network structure and the empowerment of the 3rd Gen Intel Xeon Scalable processor. Synthesis is faster without compromises on speech quality. The new solutions have proven to achieve an RTF of 0.036 in TTS[2], with a Mean Opinion Score (MOS) of 4.4.

Improved Performance - With a simplified model structure, together with linear processing, sub-band division, sparse technology and others, the custom WaveRNN vocoder model has effectively reduced computation. When used with Intel® Xeon® Scalable processors, the platform enjoys enhanced TTS performance while capable of handling higher workloads. Performance on a single processor core running 100 instances has proven to be nearly the same as one[3].

Greater Computing Power - The next-gen Intel Xeon Scalable processor's embedded hardware acceleration technology, powerful core, and larger cache have helped the Xiaowei platform obtain higher levels of performance. This has allowed the platform to serve even more enterprises and create a quality smart ecosystem that supercharges AI innovations.

AI is rolling out across industries and emerging enterprises specializing in smart products are developing solutions such as voice navigation, audiobooks, intelligent customer service, and intelligent voice input and recognition applications based on speech synthesis technology to form a complete circle of human-machine interactions. While these innovative features have made life easier, many have found these products could vary greatly in AI performances and ways of operation. The user experience is to be improved. This is largely due to the difference in platforms based on which these applications were developed. As a result, enterprises have been unable to leverage their advantages in data and technology to deliver high-quality intelligent speech services, nor create a synergy through device connectivity.

The Xiaowei intelligent speech and video service access platform was designed to overcome this bottleneck. Full-stack AI with speech and semantic capabilities are combined with Tencent Cloud services to provide users with better AI performance on the platform. With Tencent's rich portfolio and big data capability, users can also access a wide range of solutions integrating Tencent's middle platform capabilities in various scenarios. Take smart hotels. By adding Xiaowei hardware, solutions such as smart inquiries and room control are instantly endowed with rich sensory functions such as vision and hearing. Better is, these products can be linked to common mobile apps such as WeChat, WeChat Map, and WeChat Music to provide an easy and seamless experience for end users. In transportation, the platform greatly enhances user experience by tapping into the massive entertainment apps such as QQ Music and Tencent News, in addition to allowing automakers to offer smart interactions such as onboard voice-assisted navigation. Tencent's Xiaowei platform has also been widely applied in other fields such as education, finance, and media.

While working with its users to develop a robust product ecosystem, Tencent has never stopped optimizing its vocoder models by upgrading the platform's core TTS capability to take the end user experience to the next level. TTS technology can be used to convert external text inputs or computer-generated data into natural-sounding speech. It's a process that the vocoder model makes computation and analysis to output speech waveforms, of which the choice of model has a major effect on synthesis results. Traditional vocoders such as the WaveNet are deep autoregressive models based on convolutional neural networks (CNNs). The output from the previous layer is fed into the end of the input layer for convolutional iteration. The generated speech quality could approach that of a natural human voice. Nevertheless, the traditional WaveNet model is cursed with the following drawbacks in practical applications.

- First, the complex structure of WaveNet means the need for more computing power and less satisfying synthesis speed. The model may not be up to scratch in intelligent voice interaction scenarios that demand high real-time performance;
- Second, the growing prevalence of intelligent speech applications means the Xiaowei platform must support a large variety of devices. The increased workload (throughput) as a result requires vocoder models with higher TTS performance.

Therefore, Tencent was in urgent need of an advanced TTS solution capable of addressing the real-time and throughput challenges. To solve the challenges, Tencent turned to Intel, an industry leader and its long-term partner, to develop two custom TTS solutions - the pWaveNet vocoder and the WaveRNN vocoder - to further enhance the performance of its platform.

Custom Parallel WaveNet Vocoder Solution

The pWaveNet model was selected not only because of its lightweight nature, but also the introduction of probability density distillation technology to the original WaveNet model. In other words, a pre-trained WaveNet model acts as the "teacher" that guides the "student" network, which makes prediction in actual operations. The "student" network is smaller in size and is given randomized white

noise as input. It learns from the teacher's probability distribution and makes continuous adjustments to reduce variance with the "teacher" and optimize output. Where the WaveNet model relies on sequential generation, with each input sample drawn from the previous output, the pWaveNet student network learns from each of the teacher's audio sample instead of its own previous outputs. This allows for parallel computing and the generation of the entire sequence of output samples in a single pass, significantly reducing the time required for TTS.

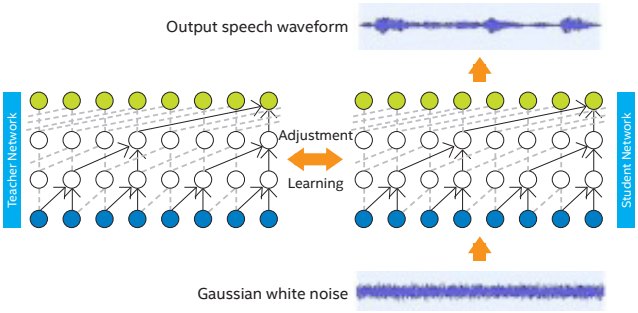

Figure 1 Structure of the Parallel WaveNet model

However, the architecture of student networks under the pWaveNet model were still based on convolutional neural network. They are smaller, but as is generally understood, convolutional operations are more computationally intensive than standard arithmetic operations. To account for this, the pWaveNet model was customized by Tencent by transforming the con1D into combination of several General Matrix Multiply (GEMM). In so doing, the network topology was simplified and computation reduced. Additionally, the OpenMP parallelism mechanism was introduced to maximize pWaveNet's advantages in parallel computing. These modifications have enabled the custom model to synthesize faster without sacrificing quality.

Custom WaveRNN Vocoder Solution

Beyond the pursuit of speed, the Xiaowei platform also faced pressure from the increasing devices connected to it, which had led to even greater demands to overall throughput. That means, in situations that require the computation of a large number of instances, each single core should service as many instances as possible. The most direct way of increasing throughput per core is to further reduce computation.

To solve this problem, Tencent chose the advanced WaveRNN model, based on which a high-performance WaveRNN TTS solution was developed. The WaveRNN is essentially a single-layer recurrent network with a dual softmax layer. The 16-bit sample sequence is divided into the coarse part (high 8-bits) and the fine part (low 8-bits). The predictions are performed by the Gated Recurrent Unit (GRU) accordingly. The single-layer recurrent network structure means only 5 computational steps are required to predict a 16-bit sample, much fewer than those required in a WaveNet DNN.

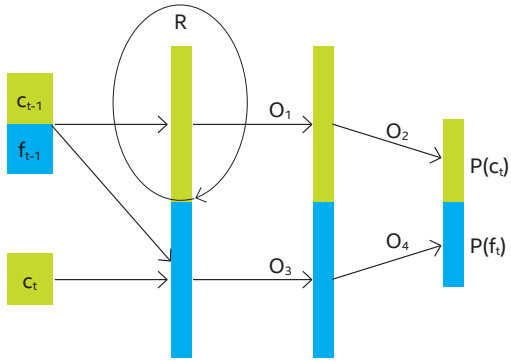

Figure 2 Structure of the WaveRNN model

Building on its inherent structural advantage, Tencent further customized the WaveRNN model to reduce computation and accelerate TTS. The sample rate network that makes the main part of the custom WaveRNN model remained a single-layer recurrent network with a dual softmax layer. Where the custom model differs is that it separates the linear part from the original input to be given prediction based on LPC, reducing computational complexity to a large extent. The sample sequence is also divided into multiple sub-bands, with computation of the subsequent sub-band beginning once the previous has been generated, effectively increasing overall computation speed. In addition, the solution uses sparse technology to reduce bandwidth demand and the overall network compute time. Large sparse models can better balance computing power in multi-core environments than small, dense models.

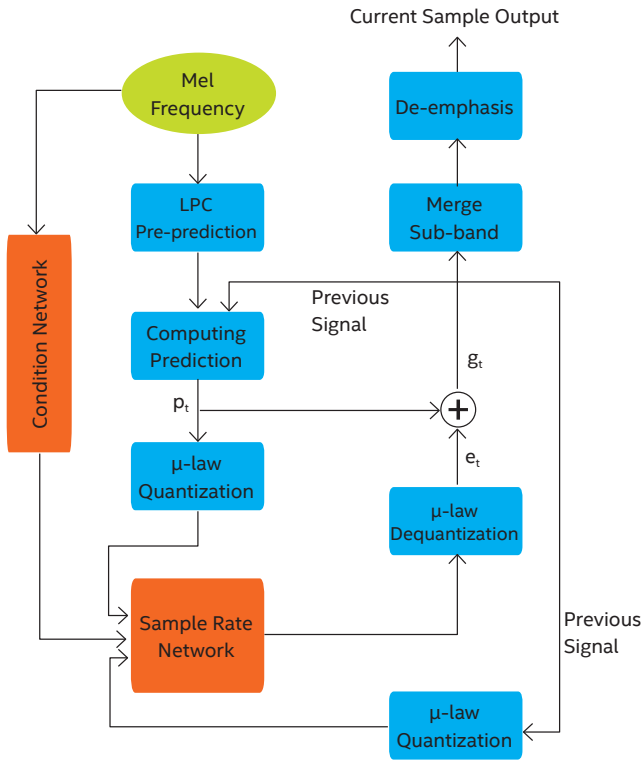

Figure 3 Structure of the custom WaveRNN vocoder model

Intel Greatly Enhances TTS Solutions

"The two keys to increasing synthesis speed are to accelerate data read/write in memory and data execution efficiency. The BF16 and Intel® AVX-512 extensions integrated in the 3rd Gen Intel® Xeon® Scalable processors have helped us achieve both goals in our custom models. With the custom pWaveNet vocoder, the platform achieved a real-time factor (RTF) of 0.036 for TTS with the quality level at MOS 4.4. The custom WaveRNN vocoder also enjoys a faster TTS speed all while handling greater workloads."

Qiao Tian
Senior Researcher
Tencent Cloud

Once the excellent model structure was decided upon, Tencent chose Intel's advanced hardware as its underlying support to maximize the performance of the entire solution. Both the custom pWaveNet vocoder model and WaveRNN model solutions utilize the 3rd Gen Intel Xeon Scalable processors. With 28 cores, the processors are capable of delivering enhanced computing power while meeting the throughput requirements of the Xiaowei platform. The embedded BF16 instructions play a critical role by effectively increasing memory utilization. When used with Intel AVX-512 instructions and the Intel oneAPI deep neural network library, the hardware can be accelerated. The new processor's extra-large cache delivers additional processing performance, which in turn improves TTS performance.

Intel® BF16 instructions reduce memory read/write times

BF16 is a new floating-point format with 1 sign bit, 8 exponent bits, and 7 mantissa bits. It can be seen as a short version of FP32 with the last 16 mantissa bits cut off. BF16 has the exact same exponent size as FP32, so it retains a similar dynamic range hence similar level of precision.

The reduction in mantissa bits, however, significantly reduces computation while improving memory storage and read/write performance. Using BF16 in the model optimization solutions resulted in FP32-level speech quality but with much shorter synthesis time.

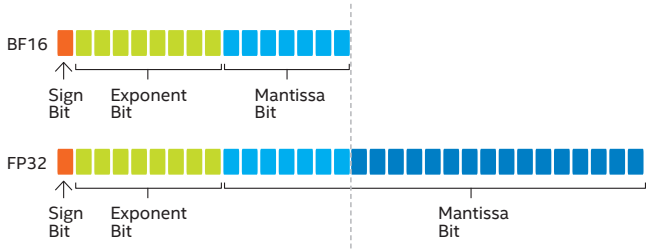

Figure 4 Structure of the BF16 and FP32 floating-point formats

Intel® AVX-512 instructions increase execution efficiency

Intel® AVX-512 is a set of instructions used for carrying out Single Instruction Multiple Data (SIMD) operations on the processor. Performance is enhanced by allowing a single processor to control multiple registers and perform data operations in parallel. Intel AVX-512 features 512-bit wide instructions to pack more operations per clock cycle. It also supports 3-Operand, with which complex, advanced instructions can be created to replace multiple simple, individual instructions to increase instruction flexibility, reduce memory access, and maximize single core execution efficiency.

Extra-large processor cache increases processing performance

Frequently accessed data is stored in the cache between the processor and memory. The processor is much faster than memory in read/write, so cache is key in providing temporary storage that is faster than memory so that the processor spends less time waiting for the data. A processor first checks the cache near it for any required data, and then memory if the data isn't found. Intel's extra-large processor cache effectively increases the cache hit rate and boost processor performance.

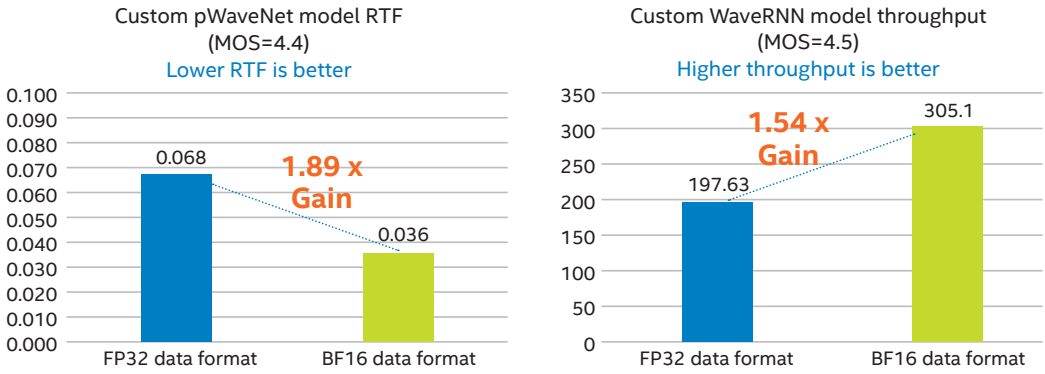

Figure 5 Increases in performance from custom solutions

Solution Performance Testing and Validation

The performance enhancements delivered by the custom solutions were jointly verified by Tencent and Intel based on the 3rd Gen Intel Xeon Scalable processors. TTS throughput and RTF were measured using BF16 and FP32 respectively to provide data support for the future expansion of Xiaowei platform.

With the same quality level (MOS 4.4), the custom pWaveNet model achieved an RTF of **0.036** and a **1.89x** performance speed-up using BF16 versus FP32[4]. The custom WaveRNN model also showed exceptional performance. It turned out that performance varied only slightly between running 1 and 100 instances on a single core. With the same quality level (MOS 4.5), total throughput reached **305.1** while a **1.54x** performance speed-up was gained using BF16 versus FP32[5].

Looking Ahead

The collaboration between Tencent and Intel has made many advanced platforms and systems a reality. With the 3rd Gen Intel® Xeon® Scalable processors, the custom solutions have delivered outstanding performance in TTS application scenarios. Next step, the two companies are planning to engage in further collaboration by integrating even more of Intel's advanced hardware and software technologies and expanding into new business scenarios. This will empower various industries to go smart by unlocking new value in speech recognition, voice print recognition, and other key AI disciplines; and step further toward a smart ecosystem where software and hardware are fully integrated.

In addition to the Xiaowei platform, Tencent and Intel will continue working together to take advantage of the excellent infrastructure offered by the next-gen Intel Xeon Scalable platform to supply users with more agile, efficient, reliable, diverse, and innovative services in cloud architecture, data cloudification, AI, high-performance computing, and security. Users will benefit from lower system administration and maintenance costs, greater agility in service deployment and launch, as well as more efforts freed to focus on business innovation to get an upper hand in the fierce market competition.

Please visit the following link for more information:
3rd Gen Intel Xeon Scalable processors: https://www.intel.com/content/www/us/en/products/docs/processors/xeon/3rd-gen-xeon-scalable-processors-brief.html
Tencent Xiaowei Platform: https://xiaowei.qcloud.com/

[1] Source: CCIC Data: Chinese Intelligent Speech Market Forecast and 2019-2021 Prospect Data: http://www.cena.com.cn/industrynews/20200109/104168.html
[2, 4] Configurations for measurements: Configurations for pWaveNet model: Configurations for FP32 solution: 3rd Gen Intel® Xeon® Scalable processor Platform, 1-node; 3rd Gen Intel® Xeon® Scalable processor CPX ES2 (QU3H), 4-Socket; 26C52T; Turbo ON; HT ON; BIOS: WCCCPX6.RPB.0018.2020.0410.1316; Memory: DDR4 2933MHz 16GB*24; Storage: Intel® SSDPE2KX010T7; NIC: Ethernet Controller 10G X550T *2; Operating System: CentOS 8.1; OS Kernel: 4.18.0-147.5.1.el8_1.x86_64; Data Analytics Acceleration Library version: 1.3; Precision: FP32; OMP_NUM_THREADS set to 1; Configurations for BF32 solution: WhiteCloudCity4S platform, 1-node; 3rd Gen Intel® Xeon® Scalable Processor CPX ES2 (QU3H), 4-Socket; 26C52T; Turbo ON; HT ON; BIOS: WCCCPX6.RPB.0018.2020.0410.1316; Memory: DDR4 2933MHz 16GB*24; Storage: Intel® SSDPE2KX010T7; NIC: Ethernet Controller 10G X550T *2; Operating System: CentOS 8.1; OS Kernel: 4.18.0-147.5.1.el8_1.x86_64; Data Analytics Acceleration Library version: 1.3; Precision: BF16; OMP_NUM_THREADSSET set to 1.
[3, 5] Configurations for WaveRNN: Configurations for FP32 solution: 3rd Gen Intel® Xeon® Scalable Processor Platform, 1-node; 3rd Gen Intel® Xeon® Scalable Processor CPX ES2 (QU3H), 4-Socket; 26C52T; Turbo ON; HT ON; BIOS: WCCCPX6.RPB.0018.2020.0410.1316; Memory: DDR4 2933MHz 16GB*24; Storage: Intel® SSDPE2KX010T7; NIC: Ethernet Controller 10G X550T *2; Operating System: CentOS 8.1; OS Kernel: 4.18.0-147.5.1.el8_1.x86_64; Data Analytics Acceleration Library version: 1.3; Precision: FP32; OMP_NUM_THREADS set to 1; Configurations for BF32 solution: WhiteCloudCity4S platform, 1-node; 3rd Gen Intel® Xeon® Scalable Processor CPX ES2 (QU3H), 4-Socket; 26C52T; Turbo ON; HT ON; BIOS: WCCCPX6.RPB.0018.2020.0410.1316; Memory: DDR4 2933MHz 16GB*24; Storage: Intel® SSDPE2KX010T7; NIC: Ethernet Controller 10G X550T *2; Operating System: CentOS 8.1; OS Kernel: 4.18.0-147.5.1.el8_1.x86_64; Data Analytics Acceleration Library version: 1.3; Precision: BF16; OMP_NUM_THREADS set to 1.