Analytical Solution of Stage-dependent Bit Resolution of Full Parallel Variable Point FFTs for Real-time DSP Implementation

Zhang, Junjie; Wang, James; Giddings, Roger; Zhang, Qianwu; Peng, Junjie; Chen, Jian; Tang, Jianming

Journal of Lightwave Technology

DOI:
10.1109/JLT.2018.2870144

Published: 15/11/2018

Peer reviewed version

Cyswllt i'r cyhoeddiad / Link to publication

Dyfniad o’r fersiwn a gyhoeddwyd / Citation for published version (APA):

Hawliau Cyffredinol / General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Analytical Solution of Stage-dependent Bit Resolution of Full Parallel Variable Point FFTs for Real-time DSP Implementation


Abstract—Digital signal processing (DSP) is a major driving force for cost-effectively realizing “software-defined anything” required by future converged networks. The fast Fourier transform (FFT) is a fundamental building block of an overwhelming majority of those DSP algorithms. For practical real-time implementation, the logic resource usage reduction of FFT operations is critical for considerably decreasing the hardware cost and power consumption. In this paper, a simple and effective solution of stage-dependent minimum bit resolution of full parallel variable-point FFTs is analytically derived, for the first time, whose validity and robustness are rigorously verified, both numerically and experimentally, over intensity modulation and direct detection (IMDD) optical OFDM transmission systems. The developed solution has unique advantages including great simplicity, excellent accuracy and robustness, and significant saving in logical resource usage. The solution can ease the practical real-time FFT DSP design, decrease the DSP complexity and maximize the overall system performance by making full use of available transceiver/system design parameters.

Index Terms—Fast Fourier transform (FFT), real-time digital signal processing (DSP), optical networks, orthogonal frequency-division multiplexing (OFDM).

I. INTRODUCTION

As a direct result of a great diversity of bandwidth-hungry services associated with newly emerging techniques such as Internet of Things (IoT) [1], 5G mobile networks [2] and latency-critical handset games, the seamless convergence of traditional optical access networks, metropolitan optical networks and mobile fronthaul/backhaul networks is regarded as a “future-proof” technical strategy to effectively address the dynamic data traffic[3][4], and also to significantly improve the signal transmission capacity and cost effectiveness. In such converged networks, it is also critical to adopt software defined networking (SDN) to enable vital networking functionalities to deliver highly desirable network operation features including, for example, flexibility, reconfigurability, elasticity, scalability and forward/backward compatibility [5]. As a major driving force of “software-defined anything”, digital signal processing (DSP) is envisaged to play a central role in practically achieving the aforementioned network operation characteristics, as DSP is capable of transparently offering, in a cost-effective manner, required network performances and networking functions [6][7].

It is well known that the fast Fourier transform/inverse FFT (FFT/IFFT) is a fundamental building block of an overwhelming majority of DSP algorithms implemented in radar imaging, audio, image, wireless local area networks (WLANs), Wi-Max, digital video broadcasting (DVB) and long term evolution (LTE)[8]-[11]. As such, the thrust of this paper is to develop, for real-time practical implementation, a simple and effective DSP solution capable of significantly reducing the FFT/IFFT DSP complexity without comprising its performance. To analytically derive the solution, for simplicity without losing any generality, throughout this paper, optical orthogonal frequency division multiplexed passive optical network (OFDM-PON) transceivers are chosen to be the special application scenario for the derived solution, since the FFT/IFFT is at the heart of those OFDM-PON transceivers [12] that inherently offer an ideal characteristic-rich environment for rigorously evaluating the solution.

Considering the fact that analogue-digital converter/digital-analogue converters (ADCs/DACs) involved in representative OFDM-PON transceivers typically operate at sampling rates of tens of GHz, the FFT/IFFT FPGA logic usage can take >80% of the total FPGA logical resources [13], thus such a huge logic usage has become one of the most significant obstacles to experimentally demonstrate real-time high-speed OFDM-PON transceivers. In addition, in typical real-time OFDM receivers, the involved FFT operation also consumes approximately 50% of DSP demodulation power [14]. The above facts indicate that for practical real-time implementation in application specific integrated circuits (ASICs), reducing the logic resource usage of the FFT/IFFT algorithm is critical for considerably...
decreasing the transceiver cost and power consumption. As such, from a practical application point of view, it is extremely valuable to explore simple and effective DSP approaches capable of considerably minimizing the FFT/IFFT FPGA logic resource usage without compromising its performance.

For real-time FFT/IFFT hardware implementation, the fixed-point arithmetic is an easy option. The finite bit resolutions in a binary format can also be adopted for both the twiddle factors and signal inputs, by taking into account the trade-off between hardware cost and FFT/IFFT operation accuracy [15]-[24].

More specifically, investigations of the impacts of FFT/IFFT bit resolution on overall OFDM transceiver performance [15] and transceiver power consumption [16] have been reported, where the 128-point FFT/IFFT operation is treated as a “black-box” without considering bit resolution variations between different intermediate FFT/IFFT operation stages. In addition, a web edition Spiral Discrete Fourier Transform (DFT)/FFT IP Core generator for FPGA hardware design has also been utilized to investigate the real-time OFDM transceiver performance [17][18]. In such a design, both the output bit resolution and the twiddle factor bit resolution are, once again, set to be intermediate FFT/IFFT operation stage-independent.

Given the fact that the DSP complexity of real-time OFDM receivers is much higher than that related to their transmitter counterparts, in this paper attention is thus focused on the receiver FFTs only. Recently, for a fixed 32-point FFT operation only, stage-dependent minimum bit resolution maps for both output bit resolution and twiddle factor bit resolution have been numerically identified, based on which minimum bit resolutions of individual DSP operations of various FFT stages can be determined by given ADC bit resolutions [19][20] in order to satisfy the overall system performance required by a specific application scenario. More recently, the above-mentioned stage-dependent minimum bit resolution maps have been significantly extended to cover the full-parallel pipelined FFTs of variable points up to 1024, and the extended bit resolution maps have also been verified experimentally [21]. It has been shown that the numerically identified maps enable significant reductions in FPGA logic resource usage without degrading the overall transceiver performance [21]. To further reduce the FFT DSP logic resource usage with the overall transceiver performance still being maintained, in [22], improved stage-dependent minimum bit resolution maps with further 3-bit reductions have been numerically identified by taking into account the DSP operation dynamic range-clipping technique, and the identified maps have also been experimentally verified for the 64-point FFT. In all of our previously published work [19]-[22], the stage-dependent minimum bit resolution maps are obtained using a numerical simulation-based sophisticated and time-consuming approach. The approach may, however, not be practically feasible for use in extremely large- and/or dynamically variable-point FFTs that are highly desirable for future converged dynamic and flexible networks. When the data traffic growth pattern is predictable, the stage-dependent directive scaling FFT operation [23] has been reported, which, according to numerically simulated results in optical OFDM systems, can tolerate occasional overflow.

In this paper, a simple and effective solution of stage-dependent minimum bit resolution of full parallel variable-point FFTs is derived analytically, for the first time, by taking into account the effects of stage-dependent clipping and input signal peak to average power ratio (PAPR). The validity and robustness of the developed analytical solution are rigorously verified, both numerically and experimentally, over intensity modulation and direct detection (IMDD) optical OFDM transmission systems subject to a wide range of various operation parameters. In comparison with our previously published work [19]-[22], the unique advantages associated with the developed analytical solution are summarized as followings:

1) Great simplicity. The solution is applicable regardless of FFT sizes, signal modulation formats and transmission system parameters. Equally important, to achieve the inverse error vector magnitude (IEVM) performance required for a given real-time transmission system, use can also be made of the solution to determine the trade-off between the allowable bit resolution and the resulting transceiver IEVM reduction.

2) Excellent accuracy and robustness. In comparisons with the ideal cases where the floating-point FFTs are adopted, the solution always gives rise to negligible IEVM differences of <0.4dB over a wide range of system operation parameters examined in the present paper. In addition, such IEVM differences are adjustable according to available hardware parameters. This feature further improves FFT operation robustness against unexpected system/network impairments.

3) Significant saving in logic resource usage. Our investigations show that >31% savings in FPGA arithmetic logic resource usage is achievable for the 128-point FFT compared with the corresponding Spiral FPGA design.

In summary, in comparison with the simulation-based sophisticated and time-consuming approach reported in [22], the analytical solution greatly eases the real-time practical FFT DSP design, considerably decreases the DSP complexity, and can serve as an effective tool for maximizing the overall system performance by making full use of available transceiver/system design parameters.

II. ANALYTICAL SOLUTION OF STAGE-DEPENDENT OUTPUT BIT RESOLUTION OF VARIABLE POINT FFTS

The Cooley-Tukey Radix-2 decimation-in-time (DIT)-based N-point FFT consists of \( \log_2 N \) stages in total. At each stage both the twiddle factor bit resolution and the output bit resolution are independently adjustable. As the search method using the bit resolution maps reported in [22] is sufficiently easy in obtaining minimum twiddle factor resolution bits, in this paper, special attention is thus given to stage-dependent output bit resolutions for the third stage and beyond, because the first and second stages just have addition and subtraction operations only.
A. Stage-dependent Quantization Noise Impact on IEVM Performance

It is well known that the N-point discrete Fourier transform (DFT) is defined as

$$X(k) = \sum_{n=0}^{N-1} x(n) \cdot W^k_n, \quad k = 0, 1, ..., N - 1$$

(1)

where $x(n)$ and $X(k)$ denote the input and output of the DFT, respectively, $n$ is the time index and $k$ is the frequency index, and $W^k_n$ is the twiddle factor defined as

$$W^k_n = e^{-j \frac{2\pi nk}{N}}$$

(2)

For an optical OFDM transceiver, $x(n)$ is real-valued and its mean value is zero. For simplicity but without losing generality, the full-scale dynamic range of $x(n)$ is assumed to be constrained to [-1, 1]. Assuming that the output signal of the $v^{th}$ FFT stage is $S_v(n)$, and let $S_0(n) = x(n)$, when a finite output resolution bit $l_{output}(v)$ is adopted for the $v^{th}$ stage only, whilst infinite output resolution bits are taken for all other remaining stages, the $v^{th}$ stage output signal with finite bit resolution $S_v(n)$ can be expressed as the sum of the corresponding output signal with infinite bit resolution $S_0(n)$ and its corresponding uniformly distributed additive quantization noise $N_v(n)$,

$$S_v(n) = S_0(n) + N_v(n)$$

(3)

The quantization noise introduced by the $v^{th}$ stage also propagates to all of the subsequent stages and affects the overall signal quality of the $N$-point FFT output. To describe the noise propagation effect, according to the Parseval’s theorem, the impact of the quantization noise imposed by the $v^{th}$ stage on the overall variance of the output signal can be expressed as

$$E[X(k)^2] = N \cdot 2^{-v} \cdot E[S_v(k)]^2$$

(4)

where $X(k)$ indicates the final frequency-domain output signal with finite bit resolution. Assuming the quantization noise associated with each individual stage is uncorrelated, by substituting Eq. (3) into Eq. (4), we have

$$E[X(k)^2] = N \cdot 2^{-v} \cdot (E[S_0(n)^2] + E[N_v(n)^2])$$

(5)

where $X(k)$ is the final frequency-domain output signal with infinite bit resolution. Eq. (5) shows that the variance of quantization noise associated with the $v^{th}$ stage increases by a factor of $N \cdot 2^{-v}$ after passing through the remaining $l_{output}(v) - v$ stages.

When a fixed digital number is used as a two’s-complement format, to prevent the signal overflow for the $v^{th}$ stage according to the Parseval’s theorem, $(v + 1)$-bits are needed for the integer part of the signal [22]. In addition, considering the fact that the output signal of the $v^{th}$ stage is complex-valued, for the signal’s real and imaginary parts, the quantization noise arising from finite output bit resolution at the $v^{th}$ stage are uniformly distributed random variables in the range of $(-\Delta/2, \Delta/2)$ with $\Delta = 2^{-l_{output}(v) - v - 1}$. As the quantization noises for both the real and imaginary parts are uncorrelated, the complex variance of $N_v(n)$ for the $v^{th}$ stage can thus be expressed as

$$E[N_v(n)^2] = 2 \cdot \int_{-\Delta/2}^{\Delta/2} y^2 dy$$

$$= \frac{2}{3} \cdot \Delta^3$$

(6)

Based on Eq. (5) and Eq. (6), the IEVM of the signal in the unit of dB subject to finite output bit resolution is given by

$$IEVM_{output, dB}(v) = 10 \cdot \log_{10} \left( \frac{E|X(k)|^2}{N \cdot 2^{-v} \cdot E[N_v(n)^2]} \right)$$

$$= 6 \cdot l_{output}(v) - 3 \cdot v + 1.76 + 10 \cdot \log_{10}(E|x(n)|^2)$$

(7)

By considering the maximum absolute value of $x(n)$ of $A_p = 1$ and the definition of the PAPR in the unit of dB at the input of the N-point FFT,

$$PAPR = -10 \cdot \log_{10}(E|x(n)|^2)$$

(8)

Eq. (7) can be rewritten as

$$IEVM_{output, dB}(v) = 6 \cdot l_{output}(v) - 3 \cdot v + 1.76 - PAPR$$

(9)

Eq. (9) indicates that for each individual stage, the IEVM performance of a signal increases by 6 dB for a 1-bit increase in output resolution bit. This feature has already been verified numerically in our previously published work [19]-[22]. It is also interesting to note in Eq. (9) that, to achieve the same IEVM performance, an about 0.5 output resolution bit increase is needed when the stage index increases by 1. Most importantly, Eq. (9) indicates that the overall IEVM performance is independent of FFT size and signal modulation format. Detailed verifications of the validity and accuracy of Eq. (9) are presented in Subsection II.D.

B. Analytical Solution of Clipping-free Stage-dependent Minimum Output Bit Resolution of N-point FFT

For a given optical OFDM PON system, the overall system IEVM performance is mainly determined by the finite bit resolution of the involved ADC/DAC, electrical-optical/optical-electrical (E/O/O-E) conversion, fiber transmission between the optical line terminal (OLT) and the optical network unit (ONU), as well as finite output bit resolution adopted in the specific FFT implementation at the receiver side [21]. When the floating IFFT operation is adopted at the transmitter side, as the stage-dependent finite output bit resolution adopted for the FFT operation is assumed to generate Gaussian noise that is uncorrelated for different stages, the overall transceiver IEVM performance, $IEVM_{total}$, can be expressed as

$$IEVM_{total} = \frac{1}{IEVM_{channel}} + \frac{1}{\sum_{v=2}^{l_{output}} l_{output}(v)}$$

(10)

where $IEVM_{channel}$ is the ideal transceiver IEVM performance when the floating-point FFT and IFFT are adopted, $IEVM_{output}(v)$ is the transceiver IEVM performance induced only by finite output bit resolutions of the $v^{th}$ FFT stage whilst the floating-point FFT is adopted for all other remaining stages.

For simplicity without losing any generality, the impact of
noise on the overall system IEVM performance can be assumed to be equal for various individual stages. Thus Eq. (10) can be simplified to
\[
\frac{1}{\text{IEVM}_{\text{total}, \text{dB}}} = \frac{1}{\text{IEVM}_{\text{channel, dB}}} + \frac{(\log_2 N) - 2}{\text{IEVM}_{\text{output}}(v)}
\]  
(11)

It should also be noted that for a specific FFT, the FFT output bit resolution of each individual stage is considered to be valid only when the following constraint is satisfied:
\[
\text{IEVM}_{\text{total}, \text{dB}} \geq \text{IEVM}_{\text{channel, dB}} - \beta
\]  
(12)
where \(\text{IEVM}_{\text{total, dB}}\) is the FFT output bit resolution of the \(N\)-point FFT operation. From Eq. (12), we can easily obtain
\[
\frac{\text{IEVM}_{\text{total, dB}}}{\text{IEVM}_{\text{channel, dB}}} \geq 10^{\frac{\beta}{10}}
\]  
(13)

By substituting Eq. (13) into Eq. (11), we have
\[
(10^{\frac{\beta}{10}} - 1) \frac{(\log_2 N) - 2}{\text{IEVM}_{\text{channel, dB}}} \geq \text{IEVM}_{\text{output}}(v)
\]  
(14)
Then
\[
\text{IEVM}_{\text{output}}(v) \geq \text{IEVM}_{\text{channel, dB}} \cdot 10^{\frac{\beta}{10}} + \gamma
\]  
(15)
where \(\gamma = 10 \cdot \log_{10} \left( \frac{\log_2 N - 2}{(\log_2 N - 2)} \right)\).

To satisfy the overall IEVM performance required for a given IMDD OFDM PON system, it can be found from Eq. (15) that the minimum IEVM performance allowed for each stage of the \(N\)-point FFT can be easily determined regardless of signal modulation format. Fig.1 shows the numerically simulated \(\gamma\) as a function of \(\beta\) for 32/64/128/256-point FFTs.

![Fig. 1. Numerically simulated \(\gamma\) as a function of \(\beta\) for 32/64/128/256-point FFTs.](image)

By considering Eq. (9) and Eq. (15), the analytical solution of the clipping-free stage-dependent minimum output bit resolution for the \(v\)-th FFT stage can be expressed as
\[
L_{\text{output}}(v) = \text{ceil} \left( \text{IEVM}_{\text{channel, dB}} + \text{PAPR} + v \cdot 3 + \gamma - 1.76 \right)
\]  
(16)
where \(\text{ceil}(\cdot)\) is the mathematical operation of rounding up to its nearest integer toward infinity.

C. Analytical Solution of Stage-dependent Minimum Output Bit Resolution of \(N\) point FFT Incorporating the Signal Clipping Effect

The key objective of this subsection is to further extend Eq. (16) by considering the stage-dependent clipping technique reported in our previous work [22]. The implementation of the clipping technique in each intermediate FFT stage ensures that the entire FFT operation dynamic range is always represented by minimum resolution bits with the negligible clipping-induced noise effect.

For a Gaussian distributed signal with a zero mean value, the variance of clipping-induced noise \(N_{\text{clip}}\) has a form given below [24]:
\[
E[|N_{\text{clip}}|^2] = 2 \int_{10^{-r}}^{10^{r}} \left( -\frac{A}{\sqrt{2\pi}} e^{-\frac{A^2}{2x^2}} + \frac{A^2}{2x^2} \right) \frac{1}{\sqrt{\pi}} e^{-y^2} dy
\]  
(17)
where \(A\) denotes the clipping value and \(\sigma\) represent the standard deviation of the Gaussian distributed signals. The SNR associated with clipping only can be expressed as:
\[
\text{SNR}_{\text{clip}} = \frac{E[|N_{\text{clip}}|^2]}{25\Omega_{\text{clip}}}
\]  
(18)
where \(\Omega_{\text{clip}} = \frac{A^2}{\sigma^2}\) denotes the clipping ratio. From Eq. (18), it can be easily seen that the clipping-induced SNR is dependent upon clipping ratio only. To explicitly demonstrate such dependence, the simulated SNR in unit dB as a function of clipping ratio in unit dB is plotted in Fig.2, which shows that a large clipping ratio results in a high SNR. However, as a large SNR normally requires high output bit resolution for achieving a targeted IEVM performance, in the IMDD OFDM PON system discussed in Section III, the clipping ratio is set at 15dB, which corresponds to a SNR as large as 90dB. This implies that a clipping ratio of 15dB is sufficiently high and its impact on the overall transceiver IEVM performance is negligible.

![Fig. 2. SNR as a function of clipping ratio for Gaussian distributed signals.](image)

After having discussed the minimum required clipping ratio for a Gaussian distributed signal, it is easier to identify the optimum stage-dependent clipped amplitude defined in Eq. (19) for the \(N\)-point FFT, which complies to approximate Gaussian distribution.
where \( A_{v,clipp} (\sigma(v)) \) is the clipped amplitude (standard deviation) of the \( v \)-th stage of the \( N \)-point FFT.

Thus the next main task is to identify the optimum stage-dependent standard deviation of the \( N \) point FFT. For the radix-2 DIT FFT architecture adopted in this paper, the output signal of the \( v \)-th FFT stage can be expressed as:

\[
S_v = S_{v-1} + W \tag{20}
\]

where \( W \) is the twiddle factor defined in Eq. (2). As the real and imaginary parts of the signal of the \((v-1)\)th stage are uncorrelated, the variance of the part and imaginary parts of the \( v \)-th stage signal can thus be written as

\[
E[S_v^2] = E[S_{v-1}^2] + E[W_v^2] \tag{21}
\]

where the suffix \( r \) represents the real part and the suffix \( i \) represents the imaginary part.

It is well known that, except the first stage where the twiddle factors only have real parts, the variances of the real and imaginary parts of the twiddle factor at each intermediate stage are:

\[
E[W_v^2] = 0.5 \tag{22}
\]

Introducing Eq. (22) into Eq. (21), we have

\[
E[S_v^2] = 1.5 \cdot E[S_{v-1}^2] + 0.5 \cdot E[S_{v-1}^2] \tag{23}
\]

As only the real part of the signal input to the FFT exists, then

\[
E[S_v^2] = E[\text{real}(x(n))^2] \cdot E[S_{v-1}^2] = 2 \cdot E[\text{real}(x(n))^2] \tag{24}
\]

Based on Eq. (23) and Eq. (24), we have

\[
E[S_v^2] = (2^{v-1} + 1) \cdot E[\text{real}(x(n))^2] \tag{25}
\]

For the output signal at each stage of the \( N \)-point FFT, Eq. (25) indicates that the variance of the real part of the signal is larger than that of the imaginary part, as such, in comparison with the imaginary part, the relatively strong clipping effect occurs for the real part. By introducing Eq. (25) into Eq. (19), the stage-dependent clipped amplitude has a form of

\[
A_{v,clipp} = \left( 2^{v-1} + 1 \right) \cdot E[\text{real}(x(n))^2]^{0.5} \cdot 10^{7.5/20} \tag{26}
\]

Given the fact that the integer part of the input signal is 1-bit, to prevent the signal overflowing, the clipping-free stage-dependent integer part is expressed as

\[
A_v = v + 1 \tag{27}
\]

By substituting Eq. (27) into Eq. (16), the stage-dependent minimum bit resolution for the fraction part can be expressed as

\[
L_{\text{output,frac}}(v) = \text{ceil} \left( \frac{1}{6} \cdot \text{IEVM}_{\text{channel,db}} + v - 1.76 + PAPR \right) - 0.5v - 1 \tag{28}
\]

On the other hand, by considering the signal clipping technique, according to Eq. (26), the stage-dependent minimum resolution bits for the integer part can be expressed as:

\[
L_{\text{output,integer}}(v) = \text{ceil} \left( \log_2 A_{v,clipp} \right) + 1 \tag{29}
\]

For a large stage index, \( 2^v + 1 = 2^v \), then Eq. (29) can be simplified to

\[
L_{\text{output,integer}}(v) = \text{ceil} \left( 0.5v - \frac{PAPR}{2} + 3 \right) \tag{30}
\]

From Eq. (28) and Eq. (30), it can be seen that a 0.5-bit decrease (increase) in output resolution bit is needed for the fractional (integer) part when FFT stage index is increased by 1. Therefore, the combination of these two aspects gives rise to almost identical output bit resolutions for every FFT stage. For a given IMDD OFDM PON system with a desired overall \( \text{IEVM}_{\text{total,db}} \), Eq. (31) determines minimum resolution bits, which provide the best trade-off between DSP complexity and \( \text{IEVM}_{\text{channel,db}} \).

D. Numerical Verifications of the Derived Analytical Solution

To numerically verify Eq. (31), use is made of an IMDD OFDM PON system, as illustrated in Fig.3, where the corresponding OFDM transceiver DSP blocks are also presented. Both the transmission system and the OFDM transceiver DSP blocks are very similar to those reported in [22].
adopted key transceiver and system parameters. The transceiver IFFT/FFT size varies from 32 to 128. The generated OFDM signal has a periodic frame structure, as illustrated in Fig. 3 (b), each frame contains a header with 80 zero-valued samples, two training sequences (TSs) each with FFT-sized samples, and 40000 data-carrying OFDM symbols. The header is used to perform coarse symbol synchronization and channel estimation. The TS generation procedure is very similar to that used in generating the data-carrying OFDM signal, except that in the TS generation, instead of an incoming PRBS, a pseudo-noise (PN) sequence is used to produce BPSK-encoded complex numbers prior to the transmitter IFFT. In each frame, cyclic prefixes (CPs) fixed at 16 samples are utilized for both the TSs and the data-carrying OFDM symbols.

In the transmitter, an incoming PRBS sequence of $2^{15}$-1 is adaptively encoded using different signal modulation formats varying among 16-QAM, 32-QAM, 64-QAM and 128-QAM. The 1st subcarrier is deactivated because of the impairments caused by low-pass filters and AC-coupling (used in our experimental platform, as discussed in Section III), whilst all other information-bearing subcarriers are arranged to satisfy the Hermitian symmetry with respect to their conjugate counterparts to generate real-valued OFDM symbols after the floating-point IFFT. After having applied 24-bit quantization and ~15 dB digital clipping for all OFDM signals regardless of the signal modulation formats and FFT sizes, the generated OFDM signals are then transferred to the OFDM receiver. In the receiver, there are major DSP functionalities including symbol synchronization, CP removal, FFT operation, channel estimation/equalization, and calculations of the IEVM/BER performance.

<table>
<thead>
<tr>
<th>Transceiver and System Parameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFT/IFFT points</td>
<td>32/64/128</td>
</tr>
<tr>
<td>Data-carrying subcarriers</td>
<td>From 2 to $2^N$-1</td>
</tr>
<tr>
<td>Modulation format</td>
<td>16/32/64/128-QAM</td>
</tr>
<tr>
<td>Cyclic prefix</td>
<td>16 samples</td>
</tr>
<tr>
<td>PRBS</td>
<td>$2^{15}$-1</td>
</tr>
</tbody>
</table>

From discussions undertaken in previous subsections, it is easy to understand that both Eq. (9) and Eq. (25) play important roles in determining the validity and accuracy of the final analytical solution presented in Eq. (31). As such, special effort should be made to verify Eq. (9) and Eq. (25). To verify Eq. (9), by making use of 32-point, 64-point and 128-point FFTs and different signal modulation formats (within an OFDM symbol, all data-carrying subcarriers encoded using the same modulation format), numerically simulated transceiver IEVM performances as a function of output bit resolution for stage 3 and beyond are presented in Fig.4(a), Fig.4(b) and Fig.4(c), where the corresponding transceiver IEVM performances calculated using Eq. (9) are also plotted for comparison. In numerically computing these three figures, floating-point IFFTs are employed in the transmitters, whilst in the receivers, except for the targeted FFT stage only, floating-point computations are applied for all other remaining FFT stages and all the twiddle factors. On the other hand, for all the considered cases of utilizing Eq. (9), fixed PAPRs of 15dB are taken.

![Simulated IEVM performances versus output bit resolution for various stages and different signal modulation formats. (a) 32-point FFT, (b) 64-point FFT, and (c) 128-point FFT.](image)

Fig. 4. Simulated IEVM performances versus output bit resolution for various stages and different signal modulation formats. (a) 32-point FFT, (b) 64-point FFT, and (c) 128-point FFT. As expected, Fig.4(a), Fig.4(b) and Fig.4(c) show that numerically simulated IEVM performances agree extremely well with the results calculated using Eq. (9), regardless of FFT sizes, signal modulation formats and FFT stage index. This indicates the validity and high accuracy of Eq. (9).

Next, to verify Eq. (25), the numerically simulated ratio between the variance of the real part, $E(\Re)$, and the variance of the imaginary part, $E(\Im)$, is plotted as a function of FFT stage index in Fig.5(a) for the 32/64/128-point FFT. Once again, in Fig. 5(a) comparisons are made between the numerically simulated results and the results predicted by Eq. (25), i.e. $2^{n+1} + 1$. In Fig. 5(a), almost perfectly overlapped curves are observed between the numerical simulations and the results predicted by Eq. (25). This confirms the validity and high accuracy of Eq. (25).

Finally, in order to numerically verify Eq. (31), additive white Gaussian noise (AWGN) is first loaded onto the encoded signals prior to the IFFT operation in the transmitter to ensure that for all the modulation formats and various FFT sizes considered, approximately 27dB IEVM performances are obtainable for the ideal cases where the floating IFFT and the floating FFT are adopted at the transmitter and the receiver, respectively. Along with the ideal IEVM performances,
IEVM performances numerically simulated using the final analytical solution in Eq. (31) are presented in Fig. 5(b), in which floating point operations are performed for the transmitter IFFT and the twiddle factors. The IEVM reduction value $\beta$ is set to 0.4dB. As seen in Fig. 5(b), $<$0.4 dB IEVM differences are achieved between the ideal case illustrated using red-dashed lines and the final solution-based results illustrated using black-dashed lines. This strongly confirms that the final analytical solution is of high accuracy.

\[ E(\text{Real})/E(\text{input}) \]

![Fig. 5 (a) Ratio between the variance of the real part and the variance of the input signal for the 32/64/128-point FFTs and 64-QAM modulation format.](image)

(b) Numerical verifications of the analytical solution for 32-point FFT, 64-point FFT, 128-point FFT and different signal modulation formats.

### III. EXPERIMENTAL VERIFICATIONS OF THE ANALYTICAL SOLUTION

The main focuses of this section are to: 1) by comparing experimentally measured overall IEVM performances between different receiver FFT designs utilizing the final analytical solution, floating point computation and the previously identified bit resolution map [22], rigorously verify the accuracy of the final analytical solution over a wide range of transceiver and system operation parameters; 2) experimentally explore the trade-off between minimum output bit resolution and the channel IEVM performance to achieve the targeted overall system IEVM performance for a specific transmission system. These two activities provide valuable guidelines for designing high performance and low DSP complexity FFTs that are highly flexible to satisfy the needs of various application scenarios.

#### A. Experimental Setup Description

To achieve the aforementioned two objectives, use is made of an IMDD OFDM experimental platform illustrated in Fig.3 (a), the key FPGA-based OFDM DSP functions and major transceiver parameters are very similar to those reported in [22]. In the OFDM transmitter, the frame structure and the DSP functions identical to those reported in Subsection II.D are implemented offline using MATLAB software, except that the 500 data-carrying OFDM symbols in the frame structure and 12-bit quantization and 12dB digital clipping are adopted in the hardware platform for 32-point FFT, 64-point FFT and 128-point FFT. As shown in Fig. 3(a), the generated OFDM signal is transferred into the internal RAM of a Xilinx ML605 FPGA board with a Virtex-6 XC6VLX240T FPGA via the UDP protocol. The ML605 FPGA board operating at 125 MHz feeds the digital OFDM signal into a 4GS/s@12-bit DAC. A narrow line-width distributed feedback laser (DFB-LD) is used to convert the electrical OFDM signal into the optical domain before injecting into a 25km standard single mode fiber (SSMF). It is also worth mentioning that the floating-point IFFT is always adopted in the OFDM transmitter side.

In the receiver side, a variable optical attenuator (VOA) is employed to adjust the received optical power. After converting the optical signal to the electronic domain by a 2.7GHz PIN, the electrical OFDM signal is amplified by a variable electrical amplifier (VEA) to ensure that the signal always occupies the entire dynamic range of a 4GS/s@10-bit ADC. 1M ADC samples are first saved using the internal on-chip RAM of another ML605 FPGA board, and then transferred back to MATLAB using the UDP protocol to perform OFDM demodulation.

#### B. Experimental Verification of the Final Analytical Solution under Fixed Transmission System Parameters

Based on the above-described experimental setup, for different FFT sizes, the experimentally measured system IEVM performances are presented in Fig. 6(a) for three different FFT designs based on the final analytical solution, the floating-point FFT and the previously reported bit resolution map [22]. In experimentally measuring Fig. 6(a), the received optical powers are fixed at -5dBm, and adaptive subcarrier bit loading is applied to maximize the achievable signal transmission capacity at overall channel BERs below the FEC limit of $3.8 \times 10^{-3}$. Fig. 6(b), Fig. 6(c) and Fig. 6(d) show the resulting subcarrier bit loading profiles and corresponding subcarrier BERs for 32-point FFT, 64-point FFT and 128-point FFT, respectively.

In addition, in experimentally measuring Fig. 6(a), in comparison with the ideal case, $\beta = 0.4dB$ is chosen for all the FFT sizes considered here. As the twiddle factor bit resolution of each individual stage imposes almost the same impact on the overall transceiver IEVM performance [22], the same twiddle factor resolution bits of 9 are considered here, which introduce an IEVM reduction as small as 0.1dB [22]. Out of the maximum allowed 0.4dB IEVM reduction, the remaining 0.3dB is contributed by the limited output bit resolution [22]. Making use of the 0.3dB IEVM performance reduction, according to the previously discussed relationship between $\gamma$ and $\beta$, the $\gamma$ values for the 32-point FFT, 64-point FFT and 128-point FFT are 16dB, 17dB and 18 dB, respectively. Finally, the experimentally optimized PAPRs of 12dB are taken for all the
Considered FFT sizes.

Compared to the floating-point-based ideal FFT operation, as expected, it is seen in Fig.6(a) that the solution-based FFT just results in <0.4dB IEVM performance reductions for all considered cases, and that good IEVM performance agreements are also observed between the solution-based FFT and the bit resolution map-based FFT. This strongly confirms the validity and high accuracy of the derived analytical solution.

To further examine the stage-dependent characteristics of the analytical solution, Fig.7 is presented, where the stage-dependent integer and fraction output bit resolutions are shown for various FFT sizes. It is shown in Fig.7 that the integer output bit resolution increases by 1-bit for every two stage increase, whilst the stage-dependent fraction output bit resolution decreases by 1-bit for every two stage increase. As a direct result, an almost identical output bit resolution of 11-bits is obtained for all the stages and FFT sizes.

C. Robustness of the Analytical Solution

To explore the robustness of the analytical solution against variations in transmission system operating parameters, Fig.8 is presented, where the IEVM/BER performances as a function of received optical power are shown for different FFT sizes: 32-point FFT in Fig.8(a), 64-point FFT in Fig.8(b) and 128-point FFT in Fig.8(c). In experimentally measuring Fig.8(a), Fig.8(b) and Fig.8(c), for each received optical power, adaptive bit loading is applied to maximize the achievable signal transmission capacity under the FEC limit of $3.8 \times 10^{-3}$, the measured signal transmission capacities are shown in Fig.8(d), where the representative constellations of individual subcarriers of the OOFDM signal corresponding to the received optical powers of -5dBm for the 64-point FFT are also inserted.

It can be seen in Fig.8 that, for various cases considered, the IEVM/BER performances obtained using the analytical solution are almost identical to those corresponding to the ideal floating-point FFT cases. More importantly, Fig.8 indicates that the validity and high accuracy of the analytical solution still remain regardless of major transceiver/system design parameters including signal modulation format, signal bit rate and received optical power. This broadens considerably the solution’s practical application range.
D. Trade-off between Bit Resolution and Channel IEVM Performance

As indicated by Eq. (28), for a specific transmission system, variations in the stage-dependent fraction bit resolution can be made dynamically to offset the channel IEVM performance in order to optimally trade between the DSP complexity and the desired overall system IEVM performance. Such a feature considerably improves system flexibility and performance robustness against unexpected system/network impairments.

The experimentally measured optimum stage-dependent output bit resolutions including the integer and fraction output bit resolution for 32/64/128-point FFTs are shown in Table II, where the conditions similar to those in Fig.7 are used whilst the respective IEVMs shown in Fig.8 are adopted for various received optical powers.

From Table II, it is seen that additional ~2.5 bits can be saved for the output bit resolution of >2 stages for the 32/64/128-point FFTs when an IEVM performance reduction of ~15dB is allowed over a wide received optical power range from -21dBm to -5dBm. The corresponding IEVM difference between the ideal case using floating-point FFTs and the considered cases in Table II for 32/64/128-point FFTs are presented in Fig.9, where <0.4 dB IEVM difference are also achieved between the ideal case and the experimental case.

<table>
<thead>
<tr>
<th>FFT SIZE</th>
<th>Received Optical Power/dBm</th>
<th>Stage Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>-5dBm</td>
<td>11 12 10 11 11</td>
</tr>
<tr>
<td></td>
<td>-9dBm</td>
<td>11 12 10 11 10</td>
</tr>
<tr>
<td></td>
<td>-13dBm</td>
<td>11 12 10 10 10</td>
</tr>
<tr>
<td></td>
<td>-17dBm</td>
<td>11 12 9 9 10</td>
</tr>
<tr>
<td></td>
<td>-21dBm</td>
<td>11 12 8 8 8</td>
</tr>
<tr>
<td>64</td>
<td>-5dBm</td>
<td>11 12 11 11 11 11</td>
</tr>
<tr>
<td></td>
<td>-9dBm</td>
<td>11 12 10 11 10 11</td>
</tr>
<tr>
<td></td>
<td>-13dBm</td>
<td>11 12 10 10 10 10</td>
</tr>
<tr>
<td></td>
<td>-17dBm</td>
<td>11 12 9 9 10 10</td>
</tr>
<tr>
<td></td>
<td>-21dBm</td>
<td>11 12 8 8 8 8</td>
</tr>
<tr>
<td>128</td>
<td>-5dBm</td>
<td>11 12 11 11 11 11 11</td>
</tr>
<tr>
<td></td>
<td>-9dBm</td>
<td>11 12 10 11 10 11 10</td>
</tr>
<tr>
<td></td>
<td>-13dBm</td>
<td>11 12 10 10 10 10 10</td>
</tr>
<tr>
<td></td>
<td>-17dBm</td>
<td>11 12 9 9 10 10 9</td>
</tr>
<tr>
<td></td>
<td>-21dBm</td>
<td>11 12 8 8 9 9 8</td>
</tr>
</tbody>
</table>

Fig. 9. IEVM difference between the ideal case using floating-point FFT and the considered cases in Table II for 32/64/128-point FFTs.
IV. Analytical Solution-associated Reduction FPGA Logic Resource Usage

For 32/64/128-point FFTs, their FPGA logic resources associated with the derived analytical solution in terms of slice registers (SR) and slice LUTs (SL) are listed in Table III, in obtaining the table, self-defined full-parallel FFTs of corresponding sizes are used. In addition, to ensure the compatibility with the last stage output bit resolution shown in Fig.6 for corresponding FFTs, the last stage output bit resolution for the Spiral FPGA design is set to 12-bit, 13-bit and 14-bits for 32-point FFT, 64-point FFT and 128-point FFT, respectively. Generally speaking, the SL usage results from arithmetic operations such as addition, subtraction and multiplication, whilst the SR usage is related to the pipelined stage FFT operations.

Table III shows that the analytical solution can save approximately 31% FPGA arithmetic logic resource usage compared with the spiral FPGA design, and 16% FPGA arithmetic logic resource usage compared with the clipping-free solution for the full-parallel 128-point FFT. It is also interesting to note that, compared with the spiral FPGA design, 13% FPGA arithmetic logic resource usage reduction can be further increased by a factor of 2 for an increase of the FFT size from 32 to 128, indicating that the analytical solution is more effective in logic resource saving for large point FFTs.

Table III

<table>
<thead>
<tr>
<th>FFT Size</th>
<th>Item</th>
<th>SR</th>
<th>SR Reduction</th>
<th>SL</th>
<th>SL Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>Spiral(12bit)</td>
<td>14288</td>
<td>24.3%</td>
<td>14643</td>
<td>4.5%</td>
</tr>
<tr>
<td></td>
<td>Clipping-free</td>
<td>11643</td>
<td>12.0%</td>
<td>15024</td>
<td>11.7%</td>
</tr>
<tr>
<td></td>
<td>Our current work</td>
<td>10811</td>
<td></td>
<td>13987</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>Spiral(13bit)</td>
<td>43153</td>
<td>20.6%</td>
<td>45785</td>
<td>17.3%</td>
</tr>
<tr>
<td></td>
<td>Clipping-free</td>
<td>37469</td>
<td>16.4%</td>
<td>41281</td>
<td>15.4%</td>
</tr>
<tr>
<td></td>
<td>Our current work</td>
<td>34269</td>
<td></td>
<td>37849</td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>Spiral(14bit)</td>
<td>117240</td>
<td>24.3%</td>
<td>142758</td>
<td>30.6%</td>
</tr>
<tr>
<td></td>
<td>Clipping-free</td>
<td>98588</td>
<td>16.1%</td>
<td>103887</td>
<td>16.1%</td>
</tr>
<tr>
<td></td>
<td>Our current work</td>
<td>88906</td>
<td></td>
<td>98947</td>
<td></td>
</tr>
</tbody>
</table>

To precisely demonstrate the analytical solution-based DSP complexity of the full-parallel N-point FFT design, the exact multipliers are listed in Table IV, which shows that 11x9/12x9-bit multiplier operations are required by the analytical solution, compared to 13x13-bit and 14x14-bit multipliers incorporated in the Spiral FPGA design for full-parallel 64-point and 128-point FFTs.

Table IV: Multipliers for Full Parallel N-point FFTs

<table>
<thead>
<tr>
<th>FFT Size</th>
<th>Multipliers</th>
<th>Total Number</th>
<th>SPIRAL Multipliers</th>
</tr>
</thead>
<tbody>
<tr>
<td>10x9</td>
<td>32</td>
<td>11x9</td>
<td>12x9</td>
</tr>
<tr>
<td>32</td>
<td>40</td>
<td>52</td>
<td>16</td>
</tr>
<tr>
<td>64</td>
<td>0</td>
<td>300</td>
<td>32</td>
</tr>
<tr>
<td>128</td>
<td>0</td>
<td>842</td>
<td>64</td>
</tr>
</tbody>
</table>

V. Conclusions

A simple and effective solution of stage-dependent minimum bit resolution of full parallel variable-point FFTs has been analytically developed, for the first time, by taking into account the effects of input signal PAPR and stage-dependent signal clipping. Extensive numerical and experimental explorations have also been undertaken to rigorously verify the validity and robustness of the developed solution over 25km SSMF IMDD optical OFDM PON systems subject to a wide range of different operation conditions. It has been shown that the solution offers up to 31% saving in FPGA arithmetic logic resource usage in comparison with the Spiral FPGA design for the full parallel 128-point FFT. The developed solution has unique advantages including great simplicity, excellent accuracy and robustness, and significant saving in FPGA logic resource usage. For practical applications, the research work has huge potential for greatly easing the real-time practical FFT DSP design, considerably decreasing the DSP complexity, and simultaneously maximizing the overall system performance by fully utilizing available transceiver/system design parameters.

REFERENCES