CONTENT
Vol. 7. No. 5-2021

A RECONFIGURABLE DATA FLOW ARCHITECTURE FOR SIGNAL PROCESSING APPLICATIONS

Amitabha Sinha, Dhruba Basu,
Department of Computer Science and Engineering Techno India,
An Engineering College affiliated to West Bengal University of Technology, Calcutta, India

ABSTRACT

This paper aims to devise a data flow model of computation for signal processing applications in which the operational nodes are signal/image processing functions such as Pixsum, Edge, Smooth [12]. These functions are configured during run time from a pool of reconfigurable FPGAS[4][5][6]. Thus because of the data flow model of computation, the signal processing functions execute concurrently. At the same time, these functions by exploiting their inherent spatial parallelism execute at high speed. There is a two fold speed up in the execution of image/signal processing applicationsone at the architecture level wherein a node of the dataflow model executes a digital signal processing (DSP) function rather than a low level machine operation. The second speed up is due to the fact that each DSP function is configured to execute in an FPGA by using maximally the concurrent operations that such a function permits. Another significant benefit that arises from our proposed architecture is that by reconfiguring an FPGA for a DSP function at run time, the reusability of the hardware elements results in reduced cost of operations. In this paper we provide an outline of the data flow architecture and its operational aspects.

KEYWORDS: *Semiconductor device modeling, RF IC Design, CMOS and BiCMOS circuit simulations, circuits for communications.*

INTRODUCTION

Intensive and complex computations are required for real time execution of signal/image processing algorithms. Such algorithms exhibit spatial parallelism and are therefore suitable for mapping into array architectures like systolic, SIMD etc. [1], [2], [3]. Speed enhancement can be achieved by using dedicated hardware that exploits the inherent parallelism in the algorithm. Data flow architecture offers a possible way of exploiting the concurrency of computations. Here firing or execution of an instruction/function is asynchronous [7], [8], [9] and depends on the availability of data from one or more processing elements (PE) which would have been fired previously. Thus the execution is very fast. However, the conventional data flow model suffers from two inefficiencies:

1 All the PEs may not be in use at any particular instant and thus the hardware utilization factor becomes low.

2. The PEs are basically ALUs and perform low level operations like Add, Sub, Mul etc. and not the functional level operations like FFT, Cosine Transforms etc..

In this paper we propose a new architecture based on the data flow principle which uses a common pool of RPEs available for the execution of image / signal processing algorithms

There is a two fold speed up in the execution of image/signal processing applications- one at the architecture level wherein a node of the dataflow model [10], [11] executes a digital signal processing (DSP) function rather than a low level machine operation. The second speed up is due to the fact that each DSP function executes in an FPGA and uses maximally the spatial parallelism that such a function possesses.

Each RPE is configured during run time to carry out a specific function as required by the algorithm. On the completion of the execution of the function, the RPE is returned to the common pool and available for being reconfigured into a new function depending on the requirement. Thus the hardware utilization is very high and the same pool of 'n' RPEs can be used again and again to execute different functions. In section II we describe the proposed architecture and the model of the RPEs that is required to support the architecture. Section III gives an overview of the state each PE can have along with the state transition diagram. The functioning of a PE manager which manages these transitions with support from the various blocks in the architecture is also discussed. We conclude in Section IV by highlighting the possible application areas.

PROPOSED ARCHITECTURE

The proposed architecture can be explained with the help of the block diagram of Figure 1. It consists of two main blocks – the Processing Unit (PU) and the Activity Template Storage Unit (ATSU). The Update Unit (UU) and the Fetch Unit (FU) act as interfaces between the PU and the ATSU.

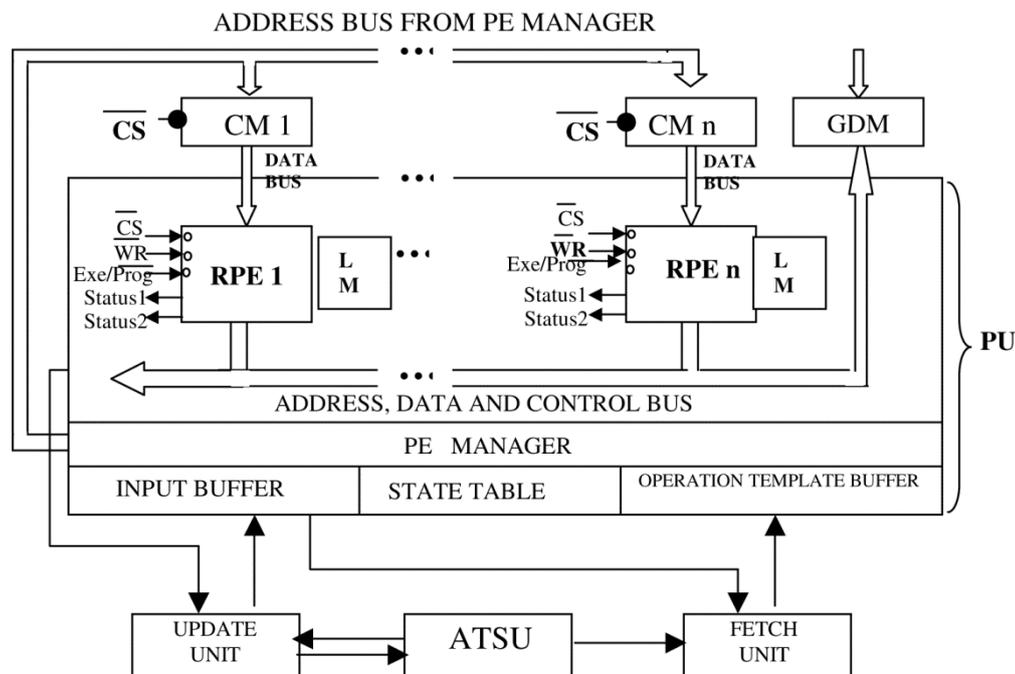The execution of the application algorithm is carried out by the n RPEs in the PU.

Figure 1

These elements are implemented using Field Programmable Gate Array(s) (FPGA) which are reconfigured for a particular function. This reconfiguration takes place by downloading the required bits-streams from Configuration Memories (CM).

Each RPE has its own local memory (LM) each of which stores all the required reconfiguration function files for the application. The LMs are connected by a global bus to a Global Data Memory (GDM). An RPE is put into execution mode when all the input operand vectors required by the function for which it has been configured are available in the GDM. This availability of input operands is as a result of the completion of execution of its predecessor functions in other RPEs according to the data flow graph. When an RPE finishes execution of a function, it stores the result vector in the GDM, outputs the result packet to the UU for transmission to the ATSU and then makes itself available for possible reconfiguration into another function as required by the data flow model of the application algorithm. The overall management of the RPE s is carried out by the PE manager in the PU. The "result packet" transmitted to the UU has the following structure.

< Result Packet > := < Result Data Pointer > <Size><{Destination Template Address in ATSU} lm >

The ATSU is a memory in which is stored information regarding the various signal/image processing functions in data structures represented by a common template (Fig. 2). In BNF notation, a template can be represented as

< Template in ATSU > := < Function> <{ operand data }1l> <{destination template address}1k > <output data pointer> <no. of elements of output data>

< operand data > := < operand data pointer > < no. of elements of operand data>

Function	
number of operand vectors (*l*)	
operand data pointer(# 1)	**T**
number of elements (# 1)	
⋮	
operand data pointer (# *l*)	**T**
number of elements (#*l*)	
number of destination templates (k)	
destination template pointer(# 1)	
⋮	
destination template pointer (# k)	
output data pointer	

Each template in the ATSU represents a distinct function in the data flow graph and its first field depicts that function. Next it has a minimum of one field to a maximum of 'l' fields of operand data, each field representing a pointer to an array in the GDM and the number of elements in each array. The template has further one field indicating the number of destination templates 'k' to which its output data should be passed

on. The addresses of these 'k' templates in the ATSU are next. The last two fields represent the starting address of the GDM into which the function output vector is written and the number of elements of the output vector respectively.
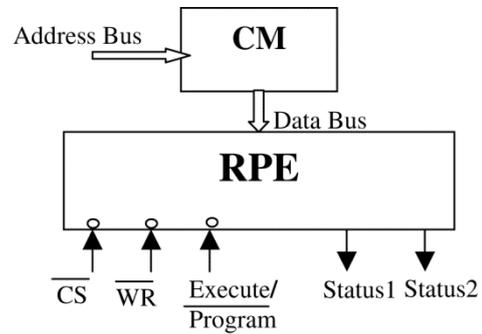

Figure 2. Model of RPE

It is clear from the above that it is the contents of the ATSU that describe the data flow model of the application algorithm to be executed in the PU. To each operand data pointer is appended a tag bit T, which are initially cleared. As and when an input operand becomes available from an executing RPE, the result packet indicates so to the UU which in turn sets the tag bit to 1. When all the tag bits of a particular template become 1, the template is ready for execution and indicates so to the UU. The UU then sends the following packet to an input buffer (IB) in the PU.

<IB Packet> : = <Function> <address of corresponding template in ATSU>.

At any point in time the IB maintains a queue of "ready to execute" functions along with the addresses of the functions' templates in the ATSU. Since the number of PEs is n, we restrict the size of the IB queue also to n as this would not affect the performance in any significant way.

It may be instructive to note that when more than n functions become ready for execution, they are kept pending in the ATSU and are transferred to the IB only when the latter's size falls below n due to the allocation of one or more functions to the PE s which have become free for allocation as determined by the PE manager. The tag bits of a template in the ATSU are cleared as soon as the function represented by the template is moved into the IB.

A state table in the PU maintains the states of the n RPEs i.e. whether these are in the running, idle or reconfiguration mode. The PE manager with the help of this state table and the IB determines whether a particular PE can be set to either execution mode or reconfiguration mode. The model of an RPE is illustrated in Fig 3 and explained below.

The address bus common to all CMs carries the addresses of the locations from which are retrieved a series of configuration bytes via dedicated data buses. The chip select bar (CS) of an RPE determines the RPE which accepts the addresses from the common address bus.

The configuration bytes are then stored internally in the RPE and help reconfigure the RPE to execute a different function. These operations on the RPE are controlled by the PE manager with the help of the signal Execute/Program. The status output signals "Status1" and "Satus2" indicate to the PE manager the status of an RPE as given in Table 1. When a function in the IB has been allotted to an RPE for execution, a request is sent by the IB to the FU to fetch the corresponding template from the ATSU. The FU obtains the template from the ATSU and puts it in the Operation Transfer Block (OTB) for formation of the result packet. Since the number of executing RPEs can be at most n, the size of the OTB should also be n.

FUNCTIONING OF THE PE MANAGER

The PE manager, which manages the n RPEs and their state transitions, consists of essentially two processes, Process 1 and Process 2. Process 1 uses the IB and a state table to manage the state transitions while Process 2 uses the output status lines status 1 and status 2 of the RPEs.

Table I

Status Outputs of an RPE

Status Output Lines		State of the RPE
Status 1	Status 2	
0	0	Idle/Execution complete
0	1	Under configuration
1	0	Configuration complete/ Ready
1	1	Executing

A state table as depicted in Table II maintains a record of the n RPEs, their current states and the functions allotted to them.

Table II

State Table i

RPE	STATE	FUNCTION
1	S_i, $i \in \{1,2,3,4,5\}$	f_j, $j \in \{1,2,\ldots z\}$ when S_i: $i \in \{2,3,4,5\}$ and f_j is undefined when S_i: $i=1$. z = number of function templates in ATSU
2	,,	,,
.	.	.
.	.	.
.	.	.
n	,,	,,

Each entry in the state table corresponds to an RPE and is identified by its RPE number. Each RPE can be in one of the five state, S_1, S_2, S_3, S_4 and S_5, where;

S_1 => yet to be configured
S_2 => under configuration
S_3 => configuration complete / ready S_4 => executing
S_5 => execution complete.

We propose that the transitions among the states in an RPE is carried out as per the state transition diagram shown in Figure 3.
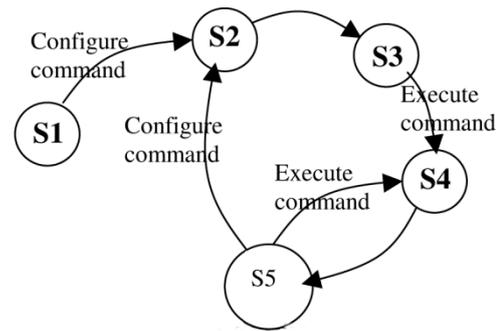

Figure 3. State Transition Diagram of an RPE

Thus initially at start up all RPEs are in S1 and make a transition to S2 as per the requirements of the algorithm. It is assumed that at start up the IB is filled with the functions which exists at the first level in the data flow graph. The corresponding entries in the state table are updated to S1.

The state table also includes the function which is under execution or under configuration in a particular RPE. It is instructive to note that when an RPE is in state 1 (as at initial start up), no function has yet been allotted to it and therefore a state table does not have a defined entry in its function column. The function that can exist in a state table can be any one of the function templates that exist in the ATSU. The IB is a first in first out (FIFO) buffer where the functions get queued in the order of their arrival from the UU and are serviced by the PE manager starting with the function at the queue head and ending with that at the queue tail according to Process 1 shown below.

Process 1 :
Do until IB queue tail
$^1\{$ take function (h) from queue head;
search state table associatively with $key_1 = <$ function(h) $>$
$< \{ S_3 | S_5 \} >$;
if at least one match exists, do the following for one of the table entries where match occurs

i) issue "Execute" command to the RPE ;
ii) update the state entry in the state table for the corresponding RPE to S_4;
iii) delete function (h) from queue and increment queue head pointer;
$\}_{11}$
else
$^{12}\{$ search state table associatively with $key_2 = < \{ S_1 | S_5 \} >$; if at least one match do the following
$^{121}\{$
i) issue "configure" (program) command to the RPE;

ii) update the corresponding state table entry to S_2;
iii) generate signals on the address bus of the CMs corresponding to function (h);
iv) activate the CS signal to the corresponding RPE;
v) increment queue head pointer;
$\}_{121}$
else
$^{122}\{$ increment queue head pointer$\}$
$\}_{122}$
$\}_{12}$
$\}_1$

It can be seen that Process 1 causes the following state transitions in Figure 4, $S_1 \rightarrow S_2$, $S_3 \rightarrow S_4$, $S_5 \rightarrow S_4$, and $S_5 \rightarrow S_2$. The transitions $S_2 \rightarrow S_3$ and $S_4 \rightarrow S_5$ require the examination of the 2 status output lines, sta-

tus 1 and status 2 for each of the n RPEs and is carried out by Process 2 as follows.

Process 2 : Do for i = 1 …….. n.
{ Read status bits : status 1, status 2,
switch (status 1i, status 2i)
{ case status 1i= 0, status 2i = 0:
update state table enter to S5; issue reconfigure command to RPE; break;
case status 1_i = 1, status 2_i = 0:
update state table entry to S2; issue execute command to RPE break;
default:
break
}

Theorem 1
Process 1 and Process 2 can run concurrently.

Proof. Process 1 reads the IB and state table and causes the following state transitions $S_1 \rightarrow S_2$, $S_3 \rightarrow S_4$, $S_5 \rightarrow S_2$, $S_5 \rightarrow S_4$ by writing into the state table and issuing either "configure" or "execute" command to the RPEs. Process 2 reads the status pins of the RPEs and causes the following state transitions $S_2 \rightarrow S_3$ and $S_4 \rightarrow S_5$ by writing into the state table. It can be seen that the two processes cause state transitions which are distinct from each other. It is shown below that logically there cannot be any conflict between the two processes when they execute concurrently. Logical conflict between the two processes can arise in either of the following two situations:

1. Two processes attempt to write two different data in the same location at the same time.
2. One process reads from while another process writes into the same location at the same time.

In our design, Process 1 writes into a location indicating states S_1, S_3 or S_5 whereas Process 2 may at the same time write into a location representing states S_2 or S_4. Therefore the two processes can never attempt to write into same location at the same time.

In the case of a simultaneous read and write, a conflict can arise if the write is a multi field write and the outcome of the read is erroneously affected if the read operation completes before the multi field write is completed. But in our proposed design the outcome of the read by Process 1 depends on the reading of one field only viz. the status field and the Process 2 also writes only one field. Therefore there is no logical conflict in the case also.

However it is possible that when Process 2 updates the state table, Process 1 tries to read it. But this problem can be solved by providing a hardware lock mechanism, which prevents such simultaneous access. This can utmost delay the execution of one of the processes for the duration of the locking but cannot cause any logical error.

CONCLUSION

In this paper we have outlined an architecture suitable for executing algorithms which possess parallelism and which need to be executed in high speed real time applications. Emerging areas such as video and image processing for 3G mobile and security applications are likely to benefit by applying our proposed architecture.

REFERENCES

[1]. S.Y.Kung. VLSI Array Processor, PrenticeHall, 1988.

[2]. H.J. Siegal, et al. PASM: partitionable SIMD/MIMD system for image processing and pattern recognition, *IEEE Trans. Computer*., Vol.C30, no. 12, pp. 934-947, Dec 1981.

[3]. A.Sinha, P.C. Jain, V. Mitra. Asynchronous SIMD – A New Architecture For A Class of Image Processing Applications, proc. *Fifth National Conf. On Communication*, pp. 287-294, IIT Kharagpur, India. 1999.

[4]. Hungwen Li and Quentin F.Stout. Reconfigurable SIMD Massively parallel Computers, *Proc. IEEE*, Vol. 79, No.4, April 1991, pp. 429-443.

[5]. John Villasenor and William H.Mangione Smith. Configurable Computing, Scientific American Article, June 1997.

[6]. John Villasenor and Brad Hutchings. The Flexibility of Configurable Computing, *IEEE Signal Processing Magazine*,1998, pp. 67-83.

[7]. J.B. Dennis. Data Flow Supercomputers, Computer, pp. 48-56, Nov., 1980.

[8]. Shuichi Sakai, Yoshinori Yamaguchi, Kei Hiraki, Yeutsu Kodama, and Toshitsugu Yuba. An Architecture of a Dataflow Single Chip Processor, *In Proc. 16th Annual Int'l Symposium on Computer Architecture*, pp 46-53. ACM, 1989.

[9]. J.B. Dennis. MIT "Stream Data Types for signal processing", Advanced topics in Dataflow Computing and Multithreading (Gao, Bic, Gaudiot eds.) IEEE Computer Society Press, 1995.

[10]. P. Netezki. Exploiting Data Parallelism in Signal Processing on a Dataflow Machine, ACM Annual Int'l Symposium on Computer Architecture, pp. 262-272, 1989.

[11].Ben Lee and A.R. Hurson. Dataflow architectures and multithreading, IEEE Computer, pp. 27-28, Aug 1994.

[12]. Edward R. Dougherty & Charles R. Giardina. Matrix Structured Image Processing, Prentice Hall, 1987.

ANALYSIS OF EMITTER COUPLED MULTIVIBRATORS

Antonio Buonomo,

Dipartimento di Ingegneria dell'Informazione, Seconda Universita degli Studi di Napoli, Napoli, Italy,
antonio.buonomo@unina2.it

Alessandro Lo Schiavo,

Dipartimento di Ingegneria dell'Informazione, Seconda Universita degli Studi di Napoli, Napoli, Italy,
loschiavo@ieee.org

ABSTRACT

The purpose of this investigation is to develop a systematic nonlinear analysis of the circuit in Figure 1, as yet not available in the literature, obtaining an exact formula for the oscillation frequency. Then, we find the condition for the circuit to behave as a relaxation oscillator, or as a nearly sinusoidal oscillator, analyzing the circuit behavior around the equilibrium point through a circuit model accounting for the transistor intrinsic capacitances. A nonlinear analysis of the popular emitter-coupled multivibrator is performed, which is based on the classical discontinuity theory. Exact relationships for calculating its waveform are obtained. Then, we investigate the circuit dynamic behavior showing that at very high frequencies the circuit exhibits a completely different behavior, similar to an LC second-order nearly sinusoidal oscillator. Conditions allowing us to predict one or the other behavior are found.

KEYWORDS: *Relaxation oscillations, emitter-coupled multivibrator, nearly sinusoidal oscillators, differential VCOs.*

INTRODUCTION

The circuit shown in Figure 1 is the well-known emitter-coupled astable multivibrator with a floating timing capacitor [1]. This circuit, and its CMOS companion [2],[3], which can be easily fabricated in monolithic form, is widely used in applications involving waveform generation, such as voltage-controlled frequency, current-controlled oscillators, and I/Q cross-coupled oscillators [4]. The circuit is commonly used as a relaxation oscillator providing a square wave at the collectors of transistors at high frequencies. However, if it is not properly designed, the circuit can exhibit a completely different behavior at very high frequency, giving rise to undesirable nearly sinusoidal oscillations whose origin is not thoroughly elucidated.

The purpose of this investigation is to develop a systematic nonlinear analysis of the circuit in Figure 1, as yet not available in the literature, obtaining an exact formula for the oscillation frequency. Then, we find the condition for the circuit to behave as a relaxation oscillator, or as a nearly sinusoidal oscillator, analyzing the circuit behavior around the equilibrium point through a circuit model accounting for the transistor intrinsic capacitances. This condition enables to elucidate the unusual dynamic behavior of the circuit. Essentially, we show that the circuit, which does not have any apparent inductor, can operate like an LC second-order oscillator, by suitably setting the circuit parameters. This is made possible as the composite two-terminal connected to the timing capacitance, which has an S-type characteristic, behaves like a locally-active inductive two-terminal, i.e. its behavior for small variations around the quiescent operating point is equivalent to a series connection of a negative resistance and an inductance.
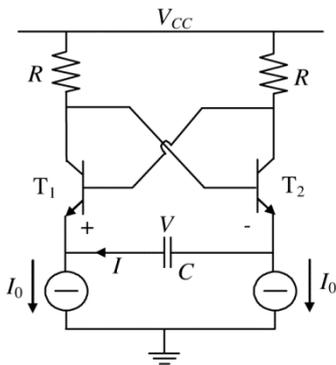

Fig. 1. Schematic of a bipolar multivibrator

Finally, circuit simulations were performed to show that both relaxation and nearly-sinusoidal oscillations can be actually generated.

BEHAVIOR AS RELAXATION OSCILLATOR

As it is known, the condition for the circuit in Figure 1 to behave as an astable multivibrator is that the *I-V* char-

acteristic of the composite two-terminal connected to the capacitance C has a negative slope at its crossing point Q with the axis $I = 0$. In fact, by virtue of the presence of C, this point defines the unique equilibrium configuration of circuit, which is instable if the characteristic slope is negative.

Assuming a purely resistive behavior of the active devices, whose collector currents are expressed by the simplified Ebers-Moll model $I_E = I_S \exp(V_{BE} / V_T)$, from equation $V = R(I_1 - I_2) + (V_{BE2} - V_{BE1})$, we deduce the nonlinear characteristic $V = F(I)$ of the composite two-terminal, which is

$$V = -2RI + V_T \ln\left(\frac{1 + I/I_0}{1 - I/I_0}\right) \tag{1}$$

This characteristic is an odd function which has the typical S-shape shown in Figure 2. To verify that the above condition is met, it is sufficient to ascertain that the derivative of (1) at the origin is negative, which happens if $V_T < RI_0$.
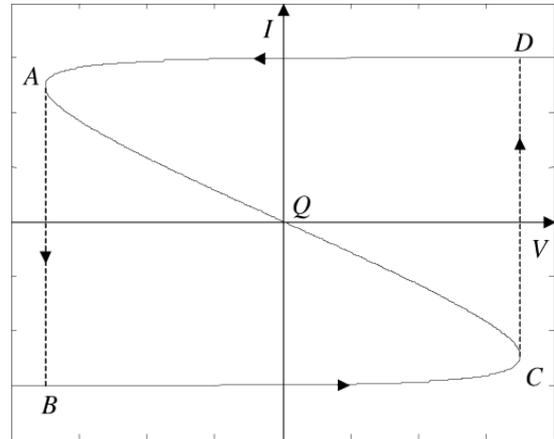
Fig. 2. Static I-V characteristic for the two-terminal connected to the capacitor

The considered circuit is a degenerated self-oscillating system wherein, as it is known, relaxation oscillations take place. Its dynamic behavior is described by the first-order nonlinear differential equation

$$I = -C\frac{d}{dt}F(I) \tag{2}$$

According to the Mandelstam postulate of constant energy [5], we assume the possibility of the existence of infinitely fast jumps of the current I. Accordingly, during the oscillatory process, the system can jump from one steady branch of the curve to the other, i.e. from the point A to point B, and from the point C to the point D. Consequently, the limit cycle is composed by two analytical branches DA and BC, which are closed by the two discontinuities AB and CD. The above assumption simplifies the calculation of the steady-state oscillation, which is here obtained by solving eq. (2).

Before solving (2), we calculate the characteristic points, $A\,(I_c, -V_c)$, $C\,(-I_c, V_c)$ (Fig. 2), which are obtained imposing that the derivative dV dI of (1) vanishes. Thus, we get the current value

$$I_c = I_0\,\sqrt{1-h}\quad,\qquad h = \frac{V_T}{R\,I_0} \tag{3}$$

and the corresponding voltage

$$V_c = -2\frac{V_T}{h}\sqrt{1-h} + V_T\ln\left(\frac{1+\sqrt{1-h}}{1-\sqrt{1-h}}\right) \tag{4}$$

The relaxation intervals, i.e. the analyticity interval of the current I, are determined solving eq. (2), which can be rearranged in the form

$$dt = \left(2RC\frac{1}{I} - \frac{2\,CI_0\,V_T}{I_0^2 - I^2}\frac{1}{I}\right)dI \tag{5}$$

Integrating this equation between the time instant immediately after a jump, t_0^+, when the current is very close to I_0 (let us call it I_i), and a time instant t, we get

$$t = t_0^+ + 2RC\,(1-h)\ln\left(\frac{I}{I_i}\right) + R\,C\,h\ln\left(\frac{I_0^2 - I^2}{I_0^2 - I_i^2}\right) \tag{6}$$

which provides a relationship between the current and the time, until the other characteristic point is reached. Therefore, we can calculate the time between two consecutive jumps, at t_0^+ and at $t_0^+ + T'$, and, thus, the period of oscillation of the multivibrator as $T = 2T'$. The oscillation frequency results in $f = 1/T$, i.e.

$$f = \frac{1}{4RC}\frac{1}{\left(1-\dfrac{V_T}{RI_0}\right)\ln\left(\dfrac{I_c}{I_i}\right) + \dfrac{V_T}{2\,RI_0}\ln\left(\dfrac{I_0^2 - I_c^2}{I_0^2 - I_i^2}\right)} \tag{7}$$

We conclude observing that the behavior analyzed is characterized by a time-domain waveform of the current presenting two intervals, during which the current varies on the S-shape characteristic from I_i to I_c , and two instantaneous jumps from I_c to I_i . Therefore, the capacitor current and, hence, the collector voltage have a near square waveform.

BEHAVIOR AS NEARLY SINUSOIDAL OSCILLATOR

Now, we show that the circuit in Figure 1 can operate either as a multivibrator, as shown in previous section, or as a nearly-sinusoidal oscillator. In particular, we perform

a small-signal analysis around the quiescent operating point Q of circuit showing that it can be unstable and that this instability can be associated with a pair of natural frequencies of the linearized equivalent circuit, which are complex-conjugate with positive real part. Consequently, nearly sinusoidal oscillations can take place in the circuit. To this end, we calculate the impedance seen looking into the emitter terminals of the circuit connected to the capacitor C, which is represented by the equivalent circuit shown in Figure 3.

This circuit is obtained using for the BJTs T_1 and T_2 the simplified small-signal high-frequency model and neglecting the output resistances.

From the circuit in Figure 3, we calculate the impedance seen looking into the emitters

$$Z(s) = k\frac{R_1 + sL_1 + s^2 d}{(1 + s\,\tau_\pi)(1 + s\,\tau_\mu)} \tag{8}$$

where

$$k = 2/(1 + g_m r_\pi)\ ; \tag{9a}$$

$$R_1 = r_b + r_\pi + R - R\,r_\pi\,g_m\ ; \tag{9b}$$

$$d = R\,r_b r_\pi C_\pi C_\mu\ ; \tag{9c}$$

$$L_1 = C_\pi r_\pi(r_b + R) + \\ + C_\mu\left[r_\pi(r_b + 4R) + r_b R(1 + g_m r_\pi)\right]\ ; \tag{9d}$$

$$\tau_\pi = \frac{C_\pi r_\pi}{(1 + g_m r_\pi)}\ ;\quad \tau_\mu = C_\mu(r_b + 4R). \tag{9e}$$
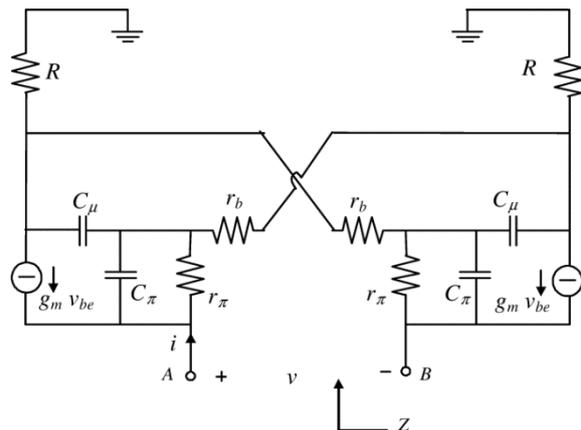
Fig. 3. Small-signal equivalent circuit of the nonlinear two-terminal connected to the capacitor

In order to investigate the behavior around the equilibrium point of the whole circuit in Figure 1, described by the parallel connection of the capacitor and the two-

terminal described by the linearized transfer function (8), we analyze the roots of the describing equation

$$Z(s)(-sC) = 1 \tag{10}$$

which can be rewritten as

$$s^3 kdC + s^2 (kL_1 C + \tau_\pi \tau_\mu)$$
$$+ s(kR_1 C + \tau_\pi + \tau_\mu) + 1 = 0 \tag{11}$$

The roots of equation (11), which can be real or complex, determine different behaviors of the circuit. Actually, if at least one root is positive then the circuit is unstable and an oscillation can take place. Moreover, if all of the roots are real the circuit behaves like a multivibrator, while if two roots are complex-conjugate the circuit behaves like a near-sinusoidal oscillator. The value of the roots and, hence, the actual behavior of the circuit depends on the values assumed by the circuit parameters, as shown in Figure 4, where the roots of (11) are plotted into the s-plane for different values of the transistor parasitic capacitances.

Parameters used are $I_0 = 3$ mA, $\beta = 50$, $R = 26\Omega$, $r_b = 160\Omega$, $C_\mu = 0,1$ C_π, $V_{CC} = 3$ V, $C = 1.5$ pF and $C\pi$ varies from 0.01 pF to 10 pF . Therefore, we deduce that increasing parasitic capacitance values make the oscillator behavior to change from relaxation to near-sinusoidal oscillator.
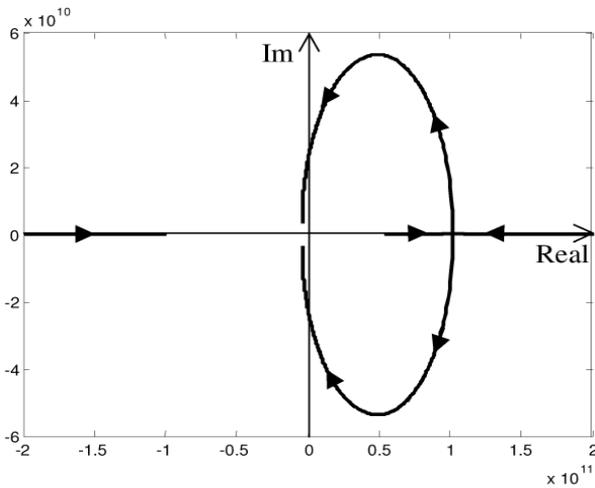
Fig. 4. Root locus for equation (11)

Let us now consider the case that eq.(11) admits a pair of complex-conjugate roots and the expression of $Z(s)$ can be simplified to

$$Z(s) = k R_1 + s k L_1 . \tag{12}$$

This impedance can be rewritten into the frequency domain in the form

$$Z(\omega) = R_e + j\omega L_e \tag{13}$$

with $R_e = kR_1$ and $L_e = kL_1$. Thus, we deduce that the behavior of the cross-coupled transistor pair at the emitter terminals can be equivalent to that of an inductance in series with a resistance, which can be made negative. Consequently, the considered two-terminal can be referred to as locally-active inductive two-terminal, and the whole circuit in Figure 1 can be represented by the series resonant circuit shown in Figure 5.
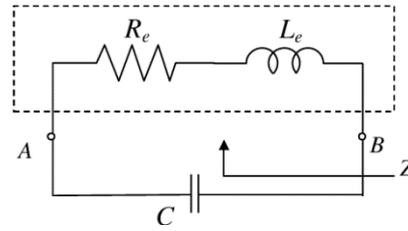
Fig. 5. Equivalent linearized circuit of the multivibrator in Figure 1

Taking into account (12), it can be shown that (10) has a pair of purely imaginary roots, $\pm j\omega_0$, with

$$\omega_0^2 = \frac{1}{LC} \; ; \qquad L = kL_1 \tag{14}$$

if the following condition holds

$$R_1 = \frac{d}{kL_1 C} . \tag{15}$$

Moreover, eq. (10) admits a pair of complex-conjugate roots with a positive real part if the following start-up condition

$$R_1 < \frac{d}{kL_1 C} \tag{16}$$

is met. Equations (14)-(16) allow us to easily design the circuit in Figure 1 in order to operate as a nearly sinusoidal oscillator. Anyway, it must be highlighted that the steady-state oscillation frequency slightly differs from that shown in (14), due to the effects of nonlinearity [6] and because the circuit parameters are set in order to satisfy (16) and not (15).

When we need to use the complete expression (8) of $Z(s)$, it is easy to show that (11) admits a pair of purely imaginary roots, $\pm j\omega_0$, with

$$\omega_0^2 = \frac{1}{LC} \; ; \qquad L = kL_1 + \frac{\tau_\pi \tau_\mu}{C} \tag{17}$$

when $R_1 = R_{lim}$, having defined

$$R_{\text{lim}} = \frac{k\,d\,C\omega_0^2 - \tau_\pi - \tau_\mu}{kC} \quad , \qquad (18)$$

and the start-up condition becomes $R_1 < R_{\text{lim}}$, which relates in a simple way the circuit behavior to circuit parameters, both active and passive.

SIMULATIONS

Numerical simulations of the considered circuit have been performed for different parameter values confirming that it can exhibit different behaviors.

Using the parameters $I_0 = 3$ mA, $\beta = 50$, $R = 26\Omega$, $r_b = 160\Omega$, $C_\pi = 0.1$ fF, $C_\mu = 0,1\ C_\pi$, $V_{CC} = 3$ V and to $C = 1.5$ pF, the roots of describing equation of the linearized model are real and two of them are positive. Therefore, the circuit is unstable and behaves like a relaxation oscillator, as shown by the SPICE transient simulation in Figure 6.
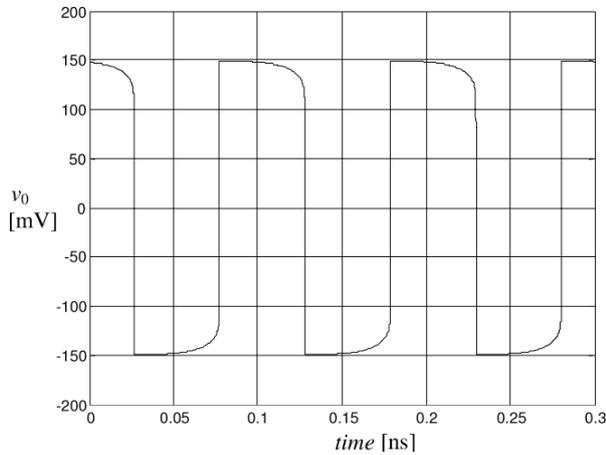
Fig. 6. Time-domain waveform produced by the circuit operating as a relaxation oscillator

On the other hand, using the same circuit parameters as above, just increasing the value of the parasitic capacitances, i.e. $C\pi = 1$ pF and $C_\mu = 0,1$ pF, the linearized model admits a pair of complex-conjugate roots with positive real part.

In this case the equivalent circuit in Figure 5 predicts an interaction between the equivalent inductance and the external capacitance, leading to a near sinusoidal oscillation. As the start-up condition is satisfied, a near sinusoidal oscillation is observed between the collectors (v_0), as shown by the SPICE transient simulation reported in Figure 7.
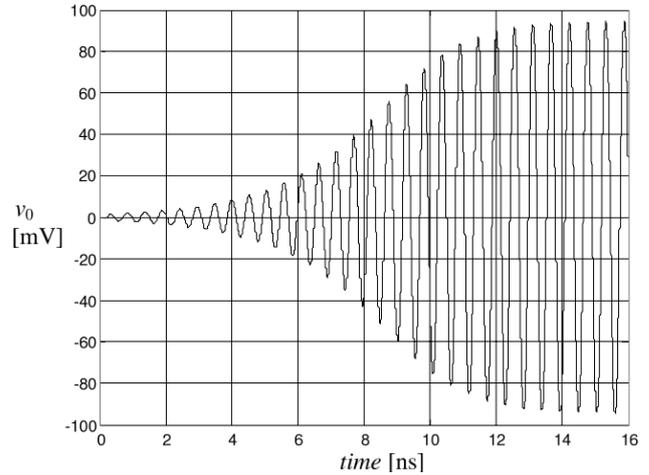
Fig. 7. Time-domain waveform produced by the circuit operating as a nearly sinusoidal oscillator

CONCLUSIONS

We presented a nonlinear analysis of the emitter coupled multivibrator, obtaining exact relationships for calculating its waveform. Then, we showed that the circuit can operate as an LC-tuned nearly-sinusoidal oscillator by suitably setting the circuit parameters. In this case, the circuit represents an alternative realization of a differential VCO which does not require an external or integrated inductor.

REFERENCES

[1] A.A. Abidi and R.G. Meyer. Noise in relaxation oscillators, *IEEE J. Solid-State Circuits*, vol. 18, pp. 794-802, Dec. 1983.

[2] I.M. Filanovsky. Remarks on design of emitter- coupled multivibrators, *IEEE Trans. on Circuits and Syst.*, vol. 35, pp. 751-755, June 1988.

[3] B.J. Song, H. Kim, Y. Choi and W. Kim. A 50% power reduction scheme for Cmos relaxation oscillator, *AP-ASIC '99*, pp. 154-157, 1999.

[4] J.R. Fernandes, M.H.L. Kouwehoven, C. van der Bos, K. van Hartingsveldt and C. Verhoeven. A 5.8 Ghz quadrature cross-coupled oscillator, *Proc. IEEE ISCAS*, vol. II, pp. 404-407, 2002.

[5] A.A. Andronov, A.A. Vitt and S.E. Khaikin, Theory of oscillators, Pergamon Press, 1966.

[6] Buonomo and A. Lo Schiavo. Analyzing the Dynamic Behavior of RF Oscillators, *IEEE Trans. on Circuits and Syst.-Part I*, vol. 49, pp. 1525-1534, 2002.

ENVELOPE FOLLOWING BASED METHOD FOR THE COMPUTATION OF LYAPUNOV EXPONENTS IN ELECTRONIC CIRCUITS

Giancarlo Storti Gajani, Amedeo Premoli, Angelo Brambilla,
Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milano, Italy
storti@elet.polimi.it

ABSTRACT

Lyapunov Exponents (LEs) can be considered as a statistical measure giving some approximate but signifficant insight in the behaviour of a dynamic system. If the systems we consider are models of realistic electronic circuits, we can reasonably assume them to be dissipative. LEs are useful to determine some general topological properties of the attractive region so that, even if the state space has a relatively large dimension, it is possible to foresee the presence of stable equilibrium points, periodic or quasi-periodic trajectories or, possibly, chaos. Lyapunov exponents and related measures such as the Lyapunov dimension are used to characterize the dynamics of complex systems and the geometrical properties of their trajectories in the state space. A few methods are available for the numerical computation of these measures and in most cases they are used only for systems that are normalized, well behaved and with a low dimensional state space. It is here proposed an approach that, being based on the envelope following method for the integration of stiff systems, can be efficiently used for the calculation of Lyapunov exponents and dimension in real electronic circuits.

KEYWORDS: *Circuit Simulation, Autonomous Circuits, Envelope Following, Lyapunov Exponents.*

INTRODUCTION

Lyapunov Exponents (LEs) can be considered as a statistical measure giving some approximate but significant insight in the behaviour of a dynamic system [1], [2]. If the systems we consider are models of realistic electronic circuits, we can reasonably assume them to be dissipative. While this assumption ensures the existence of at least one globally stable region in the state space, it does not give any information concerning the properties of this region. LEs are useful to determine some general topological properties of the attractive region so that, even if the state space has a relatively large dimension, it is possible to foresee the presence of stable equilibrium points, periodic or quasi-periodic trajectories or, possibly, chaos.

Some of these dynamic behaviours can be unforeseen by the circuit designer and, while being chaotic or even hyper-chaotic may be a desirable characteristic for some specific applications (see e.g. [3] and [4]), in many other cases chaos can have a direct impact on noise, and, in complex circuits having a state space of large dimension where hyper-chaos is a possibility, this undesirable characteristic may even be a source of noise indistinguishable from other physical sources.

Determining LEs is, unfortunately, rather expensive from a computational point of view and is only rarely performed on systems having a large dimensional state space and, possibly, very large scale differences among state variables. An interesting approach that reduces computational cost was proposed for systems with piecewise linear (PWL) nonlinearities [5], [6]. In fact trajectories of PWL systems are a succession of segments of trajectories of linear systems where each of these is determined in closed form so that the only critical problem is to determine the transition time instants between two adjacent linear regions. Since LEs can be easily computed from the system transition matrix and this is immediately available in linear systems a very fast and accurate method is obtained.

In this paper we present a new approach based on the Envelope Following (EF) integration method that, in some way, can be seen as a generalization of the previous PWL method. The EF method was originally proposed to accelerate transient analysis of stiff circuits (see e.g. [7], [8], [9]) and allows very fast convergence in the regions of state space where trajectories behave with some regularity, "slowing down" as necessary in regions where fast variations occur.

In the following we briefly recall some aspects of LEs and the PWL method for their computation. We then outline the EF method used, and, eventually, present some experimental results obtained with our own simulator, implementing the EF method and the here pro-posed LE algorithm.

LYAPUNOV EXPONENTS AND THE PWL METHOD

By assuming that state variables are well defined for the system, that it is possible to define a metric on state space and that trajectories never leave some finite subset of state space, we can say that, from a geometrical point of view, LEs correspond to a measure of how lengths, surfaces, volumes, ... are stretched or contracted in time by the system vector field. As it is well known this corresponds to a measure of how initially close trajectories will tend to become closer or to diverge versus time. All dynamic systems considered in the following are assumed, for simplicity, to be autonomous and to have one single attractive invariant set. This is not a limitation: if more attractors are present we expect the trajectories computed by EF method to remain in the proper attractive basin just as those computed by any other methods.

Consider a continuous time system

$$\dot{\mathbf{x}}(t) = \mathbf{F}(\mathbf{x}(t)) \quad \mathbf{x} \in \mathcal{R}^n \qquad (1)$$

given the initial condition $x(t_0) = x_0$ and the corresponding trajectory $\varphi(x_0, t_0, t)$ (or, for ease of notation, $\varphi(t)$), define for any $t > t_0$ the corresponding variational equation by introducing a new variable $y(t) = \varphi(t) + \Delta x(t)$ linearizing and neglecting higher order terms we have:

$$\dot{\mathbf{y}}(t) = \left. \frac{\partial \mathbf{F}}{\partial \mathbf{x}} \right|_{\mathbf{x}=\boldsymbol{\varphi}(t)} \mathbf{y}(t) \qquad (2)$$

The fundamental matrix $\mathbf{\Phi}(t)$ of the system is then defined as the general solution $y(t) = \mathbf{\Phi}(t)y(0)$ of the variational equation (2). If (1) is periodic of period T then $\mathbf{\Phi}(t + T) = \mathbf{\Phi}(t)\mathbf{\Phi}(T)$ and the eigenvalues ρ_j of $\mathbf{\Phi}(T)$ are the well known Floquet multipliers [2]. In this case LEs are easily defined as $\lambda_j \equiv (1/T) \log |\rho_j|$. Consider now a linear system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}(t) \qquad (3)$$

and its sensitivity to the Δx_0 perturbation of the initial conditions

$$\Delta \mathbf{x}(t) = e^{\mathbf{A}(t-t_0)} \Delta \mathbf{x}_0 \qquad (4)$$

choosing a norm on state space we thus have

$$\frac{\|\Delta \mathbf{x}(t)\|}{\|\Delta \mathbf{x}_0\|} = \|e^{\mathbf{A}(t-t_0)} \widetilde{\mathbf{z}}\| \qquad (5)$$

for the normalized perturbation vector $\tilde{z} = \Delta x_0 / \|\Delta x_0\|$. The right hand side of (5) is bound by a function of the leading eigenvalue of $\mathbf{A}$. In particular, for generic $\tilde{z}$ we have

$$\lim_{t \to \infty} \frac{1}{t - t_0} \log \|e^{\mathbf{A}(t-t_0)} \widetilde{\mathbf{z}}\| = \lambda_1 \qquad (6)$$

where λ_1 is the leading Lyapunov exponent. This exponent corresponds to a contraction or expansion of lengths whose effects tend to become parallel to its eigenvector. A non generic choice of $\tilde{z}$ is necessary to find contractions or expansions along other directions one for each of them.

For a non linear system such as (1) we can assume that small perturbations in the initial conditions still be-

have with an exponential evolution satisfying, in this case, the variational equation (2). We thus have for sufficiently small Δx_0 and again for a generic $\tilde{z}$:

$$\lim_{t \to \infty} \frac{1}{t} \log \|\mathbf{\Phi}(t)\tilde{\mathbf{z}}\| = \lambda_1 \tag{7}$$

and other non leading exponents can be found with specifc values of $\tilde{z}$. Direct computation of (7) for different and non generic values of $\tilde{z}$ is never performed in practice, it is preferable to compute a short step transition matrix that is orthogonalized before the effects of the leading exponent overwhelms those of lesser ones. In the case of PWL systems it is found a closed form expression for the transition matrix $\mathbf{\Phi}(t)$ that can be computed in a fast and accurate way [5], [6].

If an index is assigned to each linear subregion of the PWL system and to the corresponding transition times among these regions, we have

$$\mathbf{\Phi}(t) = e^{\mathbf{A}_k(t-t_k)}e^{\mathbf{A}_{k-1}(t_k-t_{k-1})}\cdots e^{\mathbf{A}_1(t_1-0)} \tag{8}$$

The only critical problem is the determination of transition instants; note that each "partial" transition matrix found in each linear subregion must be orthogonalized to avoid the overwhelming infuence of the leading exponent.

The method proposed in this paper can be compared in some way to the above approach. EF is used to fnd regions where, even if the system is non linear, the transition matrix can be considered constant within a given tolerance.

THE ENVELOPE FOLLOWING METHOD

Envelope Following methods have been introduced as transient analysis acceleration techniques for highly stiff oscillatory problems, see e.g. [7], [8] and [9]. Their application for the determination of steady states of oscillators has then been presented in [10]. In the sequel the basic idea on which they rely is briefy recalled. Assume that a circuit is excited by a periodic signal of period T, denoted as carrier, and that the solution can be represented as an amplitude modulated signal at least in a suitable time interval. We refer to T as the circuit working period. Suppose that modulating signal is at low frequency with respect to the carrier.

By exploiting these assumptions, we may infer that the amplitude of the solution has a small variation from one working period to the subsequent one and then a large number of periods T must be considered to have an appreciable relative variation in the amplitude of the modulated signal. This also implies that it is not needed to accurately solve the circuit in each working period but only in a reduced subset where cycles are separated by a time interval multiple of T. If this approach is adopted, the circuit may be then simulated by "sampling" the carrier at a low frequency rate, related to the modulating signal.
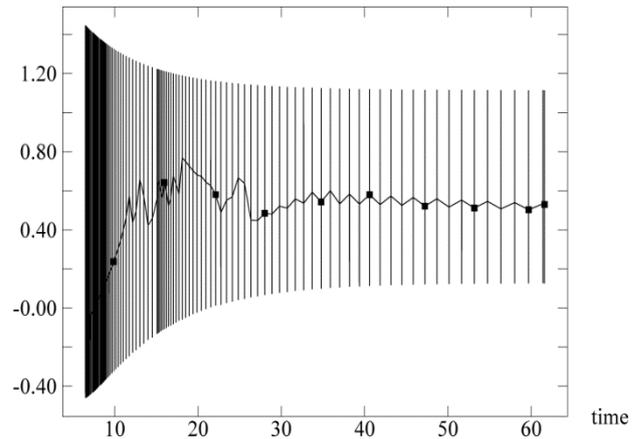
Fig. 1. Sample output waveform for the LC-tuned bipolar oscillator in Figure 2 evaluated using the EF algorithm. Time is in µs

The obtained samples constitute a discrete version of a waveform that is referred to as envelope. It is straightforward to note that the EF approach reduces the number of working periods simulated and speeds up time domain simulation. The acceleration effect is immediately visible in Figure 1, where the output voltage of the LC-tuned oscillator later used in numerical experiments is shown. From a different point of view the EF method performs a dynamic linearization of the non linear system adjusting the linearization each time a given tolerance is not satisfed. Note that the PWL approach previously outlined exploits the intrinsic characteristics of the system model, so that its transition matrix is defned as the ordered product of constant and well defned matrices, each scaled by an exponent that is proportional to the time interval in which the system remains in the corresponding region.

Even if the system model is not PWL the EF method tries to fnd regions where the transition matrix can be considered as constant within a given tolerance. At the beginning of each sample the transition matrix $\mathbf{\Phi}(T)$ corresponding to one period of length T (or approximated period if, as in general, the system is non periodic) is computed and a predictor − corrector algorithm is then used to evaluate what is the maximum time step as a multiple of the period, i.e. nT, so that the transition matrix can simply be taken to be $\mathbf{\Phi}_n(T)$. To this end consider once more (8), the corresponding equation as evaluated by the EF algorithm is thus:

$$\mathbf{\Phi}(t) = \mathbf{\Phi}_k^{n_k}(T_k)\mathbf{\Phi}_{k-1}^{n_{k-1}}(T_{k-1})\cdots\mathbf{\Phi}_1^{n_1}(T_1) \tag{9}$$

with

$$t = \sum_{j=1}^{k} n_j T_j$$

and where, for the j-th step of EF, T_j is the approx-imated period, $\mathbf{\Phi}_j(T_j)$ the corresponding transition matrix, and nj a scalar integer such that $n_j T_j$ is the maximum time interval in which the corresponding transition matrix can be

considered constant. If the system is periodic and is in its steady state $T_j = T \: \forall j$ and all partial transition matrices will be equal. In this case (9) reduces to

$$\Phi(t) = \Phi(nT) = \Phi_1^n(T) \qquad (10)$$

In our implementation we have chosen to reduce numerical problems induced by positive leading exponents by using the simple **QR** factorization method (see e.g. [1]), denoting as $\boldsymbol{Q_j R_j}$ the factorization per-formed at the j-th step, i.e. for the term $\Phi_j^{nj}(T_j)\boldsymbol{Q}_{j-1}$, we have, at the $(j + 1)$th step

$$\Phi(t) = \underbrace{\Phi_{j+1}^{n_{j+1}}(T_{j+1})\mathbf{Q}_j}_{\mathbf{Q}_{j+1}\mathbf{R}_{j+1}} \mathbf{R}_j\mathbf{R}_{j-1}\cdots\mathbf{R}_1 \qquad (11)$$

where the factorization that is performed in the cur-rent step has been evidenced. Since all R matrices are triangular, eigenvalues are immediately found as term by term products.

As an effect of the approximation introduced by EF it is interesting to note that convergence to LEs is less noisy than other methods. This is not a surprise since EF is an averaging method and tends to suppress high frequency components.

EXPERIMENTAL RESULTS

The algorithm presented in the previous sections has been implemented in our simulator [11] as an option of the EF analysis. Since EF def nes a sort of dynamic PWL approximation of a non linear system, we f rst compare our results to those obtained from the well known Chua circuit that, being PWL ab origine, is a signif cant bench-mark for our method. This circuit and the related normalization schemes are well know in literature (see [12]).

Results have been f rst compared to those outlined in [5] and [6] for the normalized Chua circuit using the reference PWL method. As it can be seen in Table I, results obtained with the two methods closely agree. The control parameter is α, and a '*' has been appended in cases corresponding to chaotic behaviour; results agree both when the dynamic behaviour is periodic or chaotic.

Table I

Lyapunov Exponents (LEs) for the normalized Chua, PWL and EF methods compared

Normalized Chua circuit						
α	PWL method			EF method		
8.0	0	-0.12	-1.22	0.0002	-0.1205	-1.2241
8.5*	0.08	0	-1.43	0.0688	0.0000	-1.4230
9.0*	0.23	0	-1.75	0.2328	-0.0112	-1.7320
9.78*	0.26	0	-2.26	0.2594	-0.0027	-2.1014
10.05	0	-0.02	-1.61	0.0001	-0.0620	-1.5695
10.5*	0.3	0	-2.25	0.2944	-0.0110	-2.1903

Results for a non-normalized Chua circuit are presented in Table II; as expected LEs have in this case a relevant scale change if compared to the corresponding normalized case (while preserving relative proportions). For example the null exponent is far from zero in absolute value, but not if compared to other exponents. In fact LEs are invariant for coordinate change, but are affected by changes in time scale (and, as expected, if these LEs are multiplied by the corresponding time scaling coefficient, in this case $\approx 7.0 \cdot 10^{-5}$, the normalized exponents are once more found). In Table II we also report the Lyapunov dimension (LD) obtained using the formula in [13], i.e. as a function of the relative values of the exponents. This measures, def ned only for strange attractors, gives some information about geometric properties of the attractor; it is also invariant for time scale changes so that, as shown in Table II, its value does not vary in the normalized or non normalized case.

Table II

Lyapunov exponents (LEs) and dimension (LD) for the non normalized Chua circuit

Non Normalized Chua circuit, EF method					
α	LEs			LD	LD (norm)
8.0	-3.1	-1728.0	-17498.5	—	—
8.5*	1183.4	-26.5	-20553.2	2.05628	2.04822
9.0*	3295.25	-144.0	-25056.7	2.1258	2.1279
10.5*	4363.39	-55.1267	-31673.7	2.13602	2.1294

An application of the method to a realistic circuit has been tested on the LC-tuned oscillator shown in Figure 2 along with circuit parameter values. The circuit has a periodic steady state behaviour as shown in Figure 1. Since all parasitic components have been included in the BJT model the state space has a larger dimension than expected by considering the transistor simply as a non linear resistive device.

Results for this circuit are summarized in Table III, the oscillator working frequency is 308.54 MHz but results have been normalized to 1 Hz.
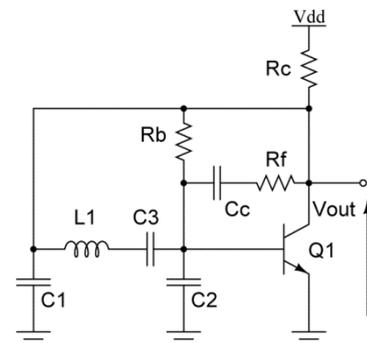

Fig. 2. The circuit of the LC-tuned bipolar oscillator: C1 = 33 pF, C2 = 33 pF, C3 = 3.17 pF, Cc = 560 pF, L1 = 100 nH, Rf = 680Ω, Rb = 100 K Ω, Rc = 1.2 K Ω

Table III

Lyapunov Exponents of the LC-tuned oscillator
in Figure 2, state space has dimension 8 for the presence
of 3 parasitic capacitances in the BJT model

LC−tuned oscillator, LEs			
0.0002	-0.0267	-0.208	-0.208
-3.27	-3.31	-3.38	-4.09

As a last example consider the Colpitts oscillator in Figure 3, the crystal is tuned at 5 MHz and has a quality factor $Q = 50000$. Since, as for the previous circuit, 3 parasitic capacitors are included in the BJT model and 3 state variables are contributed by the crystal, the state space for this circuit has a total dimension of 9.
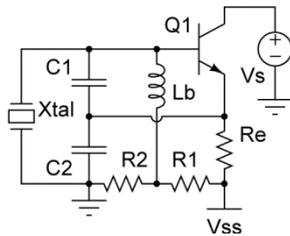

Fig. 3. The circuit of the Colpitts oscillator:
$C_1 = 33$ pF, $C_2 = 560$ pF, $L_b = 390$ µH, $R_1 = 33$ KΩ,
$R_2 = 22$ KΩ, $R_e = 1.5$ KΩ, $V_s = 10$ V, $V_{ss} = -10$ V

Also in this case the results reported in Table IV have been normalized using the oscillator frequency as scaling factor.

Table IV

Lyapunov Exponents of the Colpitts oscillator

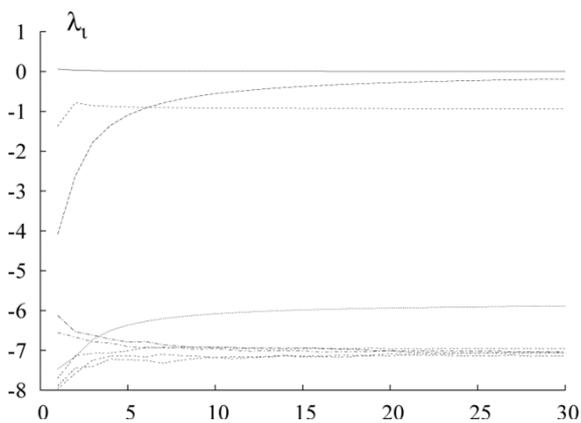| Cristall Colpitts oscillator, LEs				
0.0002	-0.0145	-0.9426	-5.7829	-6.9743
-7.0267	-7.0575	-7.0571	-7.3513	

Fig. 4. Lyapunov Exponents of the Colpitts oscillator
in Figure 3 converge in few iteration of the EF algorithm

Convergence of LEs using the proposed method is quite fast and smooth, in Figure 4 we show the first 30 iterations of the EF method. Results are within few percents of the final values already after few iterations.

CONCLUSIONS

A fast method for the determination of the complete Lyapunov spectrum of electronic circuits has been presented. The method has the advantage of computing LEs during the normal circuit simulation (with a small computational overhead). Due to averaging properties of the EF method, upon which it is based, convergence is fast and affected by a small level of numerical noise.

REFERENCES

[1] J.P. Eckmann, D. Ruelle. Ergodic theory of chaos and strange attractors, *Rev. of Modern Physics*, Vol. 57, No. 3, Part I, pp.617-656, July 1985.

[2] L.P. Shilnikov, A.L. Shilnikov, D.V. Turaev, L.O. Chua, Methods of Qualitative Theory in Nonlinear Dynamics, Part I, World Scientifc, 1998.

[3] M.P. Kennedy. Communicating with chaos: state of the art and engineering challenges, *Proc. of NDES '96, Seville*, Spain, 1996.

[4] M.Brucoli, D.Cafagna, L.Carnimeo, G.Grassi, "Secure communication via synchronization of hyperchaotic circuits", Proc. of NDES '96, Seville, Spain, 1996.

[5] M. Biey, S. Chialina, M. Hasler, A. Premoli. Piecewise-linear analysis for Chua's circuit family, including the computation of Lyapunov exponents, *Journal of Circuits, Systems and Computers,* Vol. 3, No. 2, pp. 525-536, 1993.

[6] S. Chialina, M. Hasler, A. Premoli. Fast and accurate calculation of Lyapunov exponents for piecewise-linear systems, *Int. Journal of Bifurcation and Chaos*, Vol. 4, No. 1, pp.127-136, 1994.

[7] L. Petzold. An efficient numerical method for highly oscillatory ordinary differential equations, *SIAM J. Numer. Anal.*, Vol. 18, No. 3, 1981.

[8] J. White, S.B. Leeb. An envelope following approach to switching power converter simulation, *IEEE Trans. Power Electron.,* Vol. 6, Apr. 1991.

[9] A. Brambilla, P. Maffezzoni. Envelope following method for the transient analysis of electrical circuits, *IEEE Trans. Circuits Syst.* I, Vol. 47, No. 7, pp. 999-1008, 2000.

[10] A. Brambilla, P. Maffezzoni. Envelope following method to compute steady-state solutions of electrical circuits, *IEEE Trans. Circuits Syst. I,* Vol. 50, No. 3, pp. 407-417, 2003.

[11] The Pan circuit simulator, available for Linux platforms from http://brambilla.elet.polimi.it.

[12] R.N. Madan ed., Chua's circuit: A Paradigm for Chaos, World Scientifc, 1993.

[13] P. Frederickson, J.L. Kaplan, E.D. Yorke, J.A. Yorke. The Lyapunov dimension of strange attractors, *Journ. Differential Equations*, Vol. 49, pp. 185, 1983.

LOW COST IC TESTER USING PSEUDO-RANDOM APPROACH

Liakot Ali,
Dept. of Electrical and Electronic Engineering, Universiti Putra Malaysia (UPM), Serdang, Selangor, Malaysia,
liakot@eng.upm.edu.my

Roslina Sidek,
Dept. of Electrical and Electronic Engineering, Universiti Putra Malaysia (UPM), Serdang, Selangor, Malaysia,
roslina@eng.upm.edu.my

Ishak Aris,
Dept. of Electrical and Electronic Engineering, Universiti Putra Malaysia (UPM), Serdang, Selangor, Malaysia,
ishak@eng.upm.edu.my

Mohd. Alauddin Mohd. Ali,
Dept. of Electrical, Electronic and Systems Engineering, Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor, Malaysia,
mama@eng.ukm.my

Bambang Sunaryo Suparjo,
Dept. of Electrical and Electronic Engineering, Universiti Putra Malaysia (UPM), Serdang, Selangor, Malaysia

ABSTRACT

Low cost IC testing is now a burning issue in semiconductor technology. Conventional IC tester, ATE (automatic test equipment), cannot cope with the today's continuously increasing complexities in IC technology. Deterministic algorithm, which is an idea of 1960's, is adopted in the ATE. Recently pseudo-random testing approach has been emerged as an economically viable alternative to the expensive deterministic testing. This paper introduces a SOC implementing pseudo-random test technique for low cost IC testing with reliable performance. It is capable of testing combinational circuits as well as sequential circuits with scan-path facilities efficiently. It can also be used for testing PCB interconnection faults.

KEYWORDS: *ATE, Seed, LFSR, SOC.*

INTRODUCTION

Dramatic improvement of integration technology in IC manufacturing is rapidly leading to exceedingly complex, multi-million transistor chips. All the functionalities of an electronic system are being integrated on a single chip in less than 2 cm square silicon area. This growth is expected to continue full force for the future years. With the increase of such integration densities and complexities, problems associated with testing of ICs have become much more complex and acute [1]. The cost of testing has become a major portion of the total cost of an electronic product. It is predicted in a survey that it will soon cost more to test a transistor than to make it if current trends of increasing testing cost is maintained [2]. IC testing has now become a front-end issue in semiconductor world, which needs an economic solution with reliable performance.

Modern IC manufacturing company uses ATE. The drawbacks of ATE are: (i) High equipment cost (ii) Slow test speed (iii) Huge memory requirements (iii) Tester inaccuracy. All the drawbacks of ATE are pointing towards having a new approach for cost effective IC testing. Otherwise all the benefits of semiconductor technology will be meaningless [3]. ATE is based on the deterministic algorithm. In deterministic approach [4], CUT is analyzed at the beginning of the test to generate pre-defined test patterns for detecting its faults. Deterministic test pattern enables error signals, generated due to presence of faults, and propagates them to some observable outputs from the faulty nodes or lines. This method guarantees full fault coverage but the increasing densities in the circuit leads to computational complexities and the requirements of huge amount of memory to store large test data volumes.

Pseudo-random approach [5] is now emerged as an economically viable alternative to the expensive deterministic testing. In this approach, a set of test vectors is generated randomly from 2^n possible input patterns (n= number of inputs). Linear feedback shift registers (LFSRs) are commonly used for test pattern generation because it has simple structure and can also be used as output response analyzer. The main advantage of this approach is that random pattern generation circuitry is simple and a large number of test pattern can be generated using smaller data storage.

The attempt to design and develop low cost IC tester is prevalent among the researchers since 1970. A small tester was built in the Laboratoire d'Automatique de Grenoble in 1970 and performed some random testing trials [6]. Test processor chip based on weighted random test techniques have been reported [7]. This paper introduces a SOC using pseudo-random test technique to develop a cost effective IC tester. The tester needs simpler hardware and control complexities and less memory space than that of others. Tester inaccuracy problem has also been addressed by introducing the concept of single chip IC tester which has been designed using industry standard HDL.

In order to maintain the accuracy of the tester, it is essential that the fabrication technology of the tester circuit should be at the leading edge with respect to the CUT. Since the design of the proposed SOC is in Verilog HDL, which is technology independent, the soft core can be reused to new fabrication technology keeping pace with the today's continuously changing technology environment. The performance of the IC tester has been verified using fault simulation technique. The following sections of this paper describe the design and operation of the SOC and fault simulation results of benchmark circuits.

DESIGN OF THE SOC

Figure 1 shows the functional block diagram of the SOC. Main modules of the SOC are a micro-UART, controller unit, test pattern generator (PG), buffer register (BR), signature analyzer (SA) and memory modules (RAMs). All the modules except memory modules have been designed using verilog HDL. The memory modules have been used from Altera Quartus library. The design has been compiled and simulated using the Qurtus EDA tool.
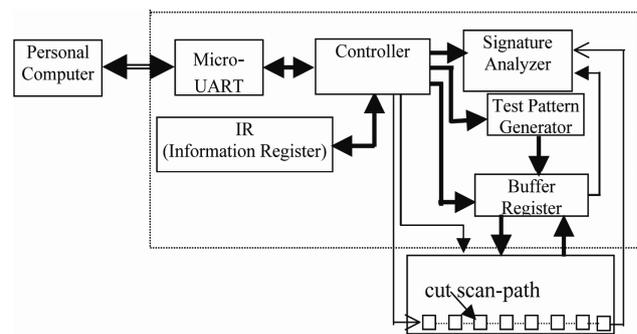
Fig. 1. Functional block diagram of the IC tester

DESCRIPTION OF THE MODULES

Micro-UART: It consists of a transmitter and a receiver module. Data communication between personal computer (PC) and the SOC is performed through the UART. It is named as micro-UART due to its modularity, configurability and extremely compact size [11]. A user friendly GUI as shown in the Figure 2 has been developed using C++ Builder for key-in test information such as numer of primary I/O, number of CUT scan-path, number of test set etc.

Controller: It consists of finite state machines (FSMs) and datapaths. FSMs generate sequence of necessary controlling pulses for loading data in the memory and executing the test process of the CUT.

Memory and Information Register (IR): Necessary information for testing CUT such as number of primary inputs, primary outputs, test set, test length etc. is stored in the IR and memory modules (RAMs). Three RAMs named as test length storage RAM (RAM_TL), seed stor-

age RAM (RAM_SD) for random pattern generation and signature storage RAM (RAM_SG) have been used.

Pattern Generator (PG): It is a 32 bit LFSR and its feedback connection is according to primitive polynomial $P(X) = 1 + X + X^{27} + X^{28} + X^{32}$.

Buffer register (BR): The BR is also 32 bit. Test pattern generated from the PG is loaded into the BR and is applied to the CUT and the output response is also captured into the BR and is sent to the SA.

Signature Analyzer (SA): The SA has similar structure of the PG except having two inputs. The state of the SA is given by the equation:

$S(t+1) = [T] . [S(t)]^T + [R(t)]^T$ where, $S(t+1)$ and $S(t)$ are SA states at times t+1 and t respectively. [T] and $[R(t)]^T$ are the transition matrix of the SA and are the transition matrix of the SA and Aliasing error of the SA is $1/2^{32}$.

Fig. 2. Graphical user interface (GUI)
between the SOC and the PC

OPERATION OF THE SOC

Operation of the SOC has two phases: (a) load/write data in the IR and the memory modules (b) circuit test and retrieval of test result. These two operations are controlled from the GUI as shown in the Figure 2.

Load/Write: Prior to IC testing, the following information is loaded in the IR and memory modules.

- Load the IR with necessary information such as number of CUT primary inputs and outputs, number of CUT scan paths, number of test sets,
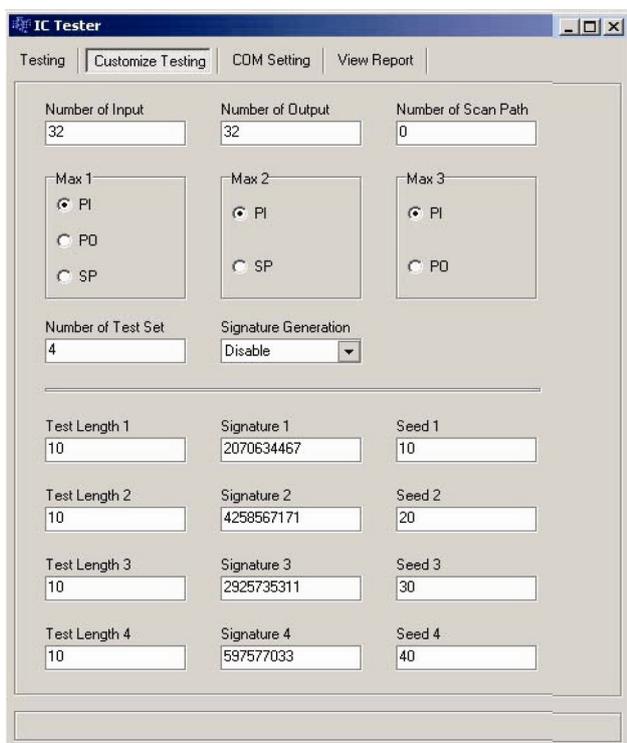
- Load the test length of each test set in the test length storage RAM
- Load the data in the seed storage RAM for random test pattern generation
- Load the reference signature in the signature storage RAM.

Circuit test and response retrieval: At the beginning of test of IC, the PG, BR and SA are reset to zero and the test mode is set on from the GUI. The test sequences are as follows:

```
---------------------------------------------------------------
Loop1
   If test mode on {
   Read the test length for the 1st test set,
   Read the seed value of LFSR for the 1st test set,
   Read the reference signature for the 1st test set,
   Set seed in LFSR by setting input set_seed=1,
   Loop2
   Enable LFSR for max(pi,sp) clock cycles,
   Enable buffer register for max (pi,po) clock cycles,
   Enable signature analyzer for max(pi,po,sp) clock
   cycles,
   Set the CUT to normal mode by setting input
   cnm=1 ;
   If cnm=1 {
      Apply test vector to the CUT
      Capture response and send to signature analyzer
      Increment test_counter value}
   If test_counter value=presetted test length {
         End_set(e_set)=1 ;
         Compare signature with reference signature }
Else goto Loop2
If e_set=1 {
         Increment set_counter, }
If set counter_value=presetted number of test set {
         End_test (e_test)=1,
         Ready for next test, }
}
Else goto Loop 1
---------------------------------------------------------
```

SPECIFICATION OF THE SOC

Specifications of the SOC are as follows :
- Maximum CUT input/output pins: 32
- Maximum CUT scan-path: 128
- Maximum test: 4
- Maximum generated test vectors: 1.7×10^7
- Maximum test length 2^{14} K

FAULT SIMULATION RESULTS

For the evaluation of the performance of the IC tester, fault simualtion experiments on ISCAS85 benchmark circuits [9] have been conducted using TetraMax fault simulator of the Synopsis EDA tool. Pseudo-random testing approach has the drawbacks of requirements of higher number of test vectors to achieve acceptable fault coverage. This problem has been overcome by using appropriate seed of the LFSR.

Summary of the fault simulation results have been presented in the Table 1. It shows that if the LFSR is initialized using appropriate seed then it is capable of producing acceptable fault coverage using lower number of test vectors. The results are comparable with that of other researchers [7, 10].

Table 1

Summary of the fault simulation results

ISCAS85 Benchmark Circuits	Number of test vector	% Fault coverage
C432	288	100
C499	928	100
C880	384	98.74
C1355	1300	100
C1908	2208	100
C5315	1408	99.13
C6288	128	100

CONCLUSIONS

Design of a cost effective IC tester has been presented. Since the IC tester is SOC, it posses improved performance, improved reliability, reduced power consumption, reduced cost, reduced system size and short time-to-market. It requires lower data storage requirements. The performance of the tester verified and showed that it is capable of producing 100% fault coverage using lower number of test vectors.

ACKNOWLEDGMENT

REFERENCES

[1] A. Liakot, S. Roslina, A. Ishak, M.A. Alauddin, S.S. Bambang. Challenges and Directions for IC testing, *Integration, the VLSI Journal,* pp. 17-28, Vol 37(1), Elsevier Science, Netherland.

[2] W.M. Needham. Nanometer technology challenges for test and test equipment, *Comput.,* 32 (11) (1999), pp. 52-57.

[3] V.D. Agrawal, C.S. Sharad. Test Generation for VLSI Chips, *Comp. Society Press*, Washington, 1988.

[4] I. Hamzaoglu, J.H. Patel. New techniques for deterministic test pattern generation, *16th IEEE Proceedings of VLSI Test Symposium,* (1998), pp. 446-452.

[5] P.H. Bardell, W.H. McAnney, J. Savir. Built-in Test for VLSI: Pseudorandom Techniques, *John Wiley and Sons,* New York, USA, 1987.

[6] R. David. Random Testing of Digital Circuits: Theory and Applications, Marcel Dekker Inc., New York, 1998.

[7] A. Iftekhar. VLSI circuit testing using probabilistic approach, Ph.D. Thesis, Universiti Kebangsaan Malaysia, 1995.

[8] A. Liakot, S. Roslina, A. Ishak, M.A. Alauddin, S.S. Bambang. Design of a Micro-UART for SoC application, Int. Journal of Comp. and Electrical Engineering, Elsevier Science, USA, 2004.

[9] F. Braglez, H. Fuziwara. A neural netlist of 10 combinational bench-mark circuits and target translator in FORTRAN, *Special session on ATPG and fault simulation, International Symposium on Circuits and Systems*, Kyoto, Japan, June, 1985.

[10] H.-J. Wunderlich. Multiple distributions for biased random test patterns, *IEEE Trans. on Comp.-Aided Design*, Vol. 9 (6), pp. 584-593, 1990.

PROPAGATION EFFECTS ON Z PARAMETERS IN AN FET EQUIVALENT CIRCUIT

M. Balti,
ENIT, Tunis, Tunisia

D. Pasquet,
ENSEA/ECIME, Cergy, France,
pasquet@ensea.fr

M. Ariaudo,
ENSEA/ECIME, Cergy, France

A. Samet,
EPT, La Marsa, Tunisia

ABSTRACT

The design of microwave circuits needs a good analysis of the performances of the field-effect transistor equivalent circuit. Indeed the small signal equivalent circuit of the field-effect transistors makes it possible to easily determine their performances such as the gain and the noise figure. A field-effect transistor constitutes a propagation structure along its gate width. Telegraphists' equations are solved for this structure. One deduces from this the effect of the propagation on the transistor Z-parameters which can be taken into account in electric simulations and which may improve the use of long transistors at lower frequencies and of short transistors at higher frequencies [1].

KEYWORDS: *FET equivalent circuit, field-effect transistor,* microwave circuits, *Z-parameters.*

The article is reworked from unpublished 2nd IEEE International Conference on Circuits and Systems for Communications (ICCSC) materials.

INTRODUCTION

The small signal equivalent circuit of the field-effect transistors makes it possible to easily determine their performances such as the gain and the noise figure. However this conventional equivalent circuit shown in Figure 1 does not take into account all the effects of propagation which can occur along its structure due to the distributed nature of the gate.
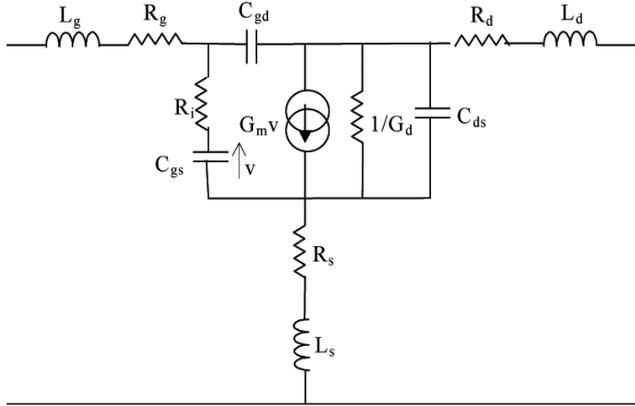

Fig. 1. Equivalent diagram of a field-effect transistor

It is convenient to use the admittance parameters to characterize the transistor electrical properties, which were given in system (1).

$$\begin{cases} y'_{11} = R_i C_{gs}{}^2 \omega_2 + j\omega(C_{gs} + C_{gd}) \\ y'_{12} = -j\omega C_{gd} \\ y'_{21} = g_m - j\omega(C_{gd} + g_m(R_i C_{gs} + \tau)) \\ y'_{22} = g_d + j\omega(C_{ds} + C_{gd}) \end{cases} \tag{1}$$

Several works showed the propagation effects on frequency limitation of the transistor [2, 3, 4]. Other works proposed structures not only to overcome wave propagation losses on gate electrode but also to take advantage of travelling wave effects [6, 7, 8, 9].

To overcome this difficulty, we developed a new method allowing to obtain a new equivalent circuit of the field-effect transistors which takes into account the propagation phenomena. In section II of this paper we describe our new method by detailing the various theoretical stages. The variation of the device [Z] parameters according to the gate width are shown in section III. We finally show the propagation phenomena in terms of new elements of the equivalent circuit.

THEORETICAL ANALYSIS

Figure 2 indicates the zone dealt with in this paper.

The three metallic pads giving access to the active part of the transistor are located apart from the rectangular frame. These pads are related to the three electrodes of the transistor which are the gate, the source and the drain. They can be simulated by an 2.5D electromagnetic tool.
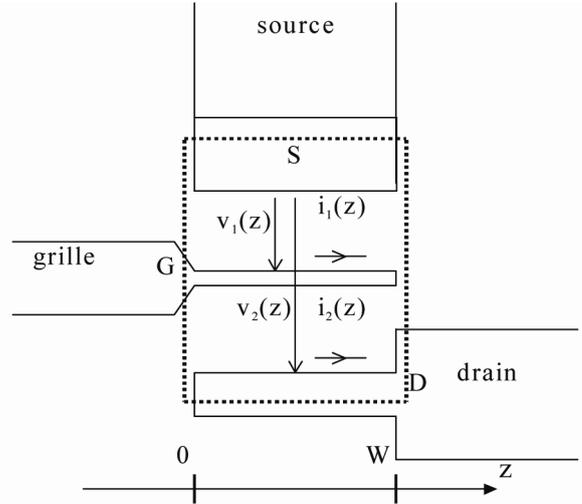
Fig. 2. Equivalent circuit of a field-effect transistor

The part within the rectangular frame is regarded as a parallel lines structure whose elementary section length is modelled by series parameters corresponding to the currents parallel to the gate, and parallel parameters corresponding to the currents perpendicular to the gate. The source current is perpendicular to the gate. So the access resistance r_s is included into the parallel elements and the self and mutual inductances are neglected. The gate current is parallel to the gate, so the access resistance r_1 and inductance l_1 are included into the series elements. The drain current includes both parallel and perpendicular components. So the access resistance consists of a series r_2 and a parallel rd elements, and the access inductance l_2 is a series element.

Figure 3 represents in a simplified way, a δx long field-effect transistor section.
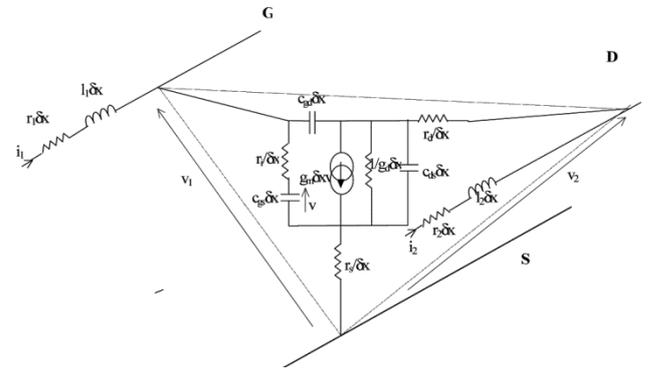
Fig. 3. A δx long section field-effect transistor section

The field-effect transistor extrinsic series parts constitute a propagation structure whose primary parameters are r_1, l_1, r_2 and l_2 .

The transistor intrinsic part of width δx may be described by its admittance matrix $[y']\delta x$. The parallel access elements r_s/δ_X and r_d/δ_X are put in series respectively with the source and the drain. They all are described by the admittance matrix $[y]\delta x$.

It is assumed that, in the gate access plane ($x=0$) the voltage v_1 and the current i_1 represent respectively the gate voltage and current, while the open end of the drain is open. So the current i_2 is zero.

At the other end, in the drain access plane ($x=W$), the voltage v_2 and the current $-i_2$ represents respectively the drain voltage and current, while the current $i1$ at the open end of the gate is zero.

The resolution of the telegraphists'equations with these boundary conditions expressed in (2) allows us to obtain the complete electric parameters of the transistor in terms of a complicated combination of hyperbolic functions (3)

$$\begin{cases} i_1(W)=i_2(0)=0 \quad \text{open ended line} \\ v_1(0)=v_{gs} \\ v_2(W)=v_{ds} \end{cases} \tag{2}$$

$$\begin{bmatrix} Z_{11}=\dfrac{z_1}{\sqrt{(A+B)^2-4(AD-BC)}}\left[-\dfrac{\gamma_2^2-A}{\gamma_1}COth\,\gamma_1 W+\dfrac{\gamma_1^2-A}{\gamma_2}COth\,\gamma_2 W\right] \\[4mm] Z_{21}=-\dfrac{z_1 C}{\gamma_1^2-\gamma_2^2}\left(\dfrac{1}{\gamma_2 Sh\gamma_2 W}-\dfrac{1}{\gamma_1 Sh\gamma_1 W}\right) \\[4mm] Z_{12}=\dfrac{B z_2}{\sqrt{(A+D)^2-4(AD-BC)}}\left[\dfrac{1}{\gamma_1 Sh\gamma_1 W}-\dfrac{1}{\gamma_2 Sh\gamma_2 W}\right] \\[4mm] Z_{22}=\dfrac{z_2}{\sqrt{(A+D)^2-4(AD-BC)}}\left[\dfrac{\gamma_2^2-A}{\gamma_2}COth\,\gamma_2-\dfrac{\gamma_1^2-A}{\gamma_1}COth\,\gamma_1 W\right] \end{bmatrix} \tag{3}$$

Two propagation modes appear with $\gamma 1$ et $\gamma 2$ as propagation constants [5]: a slow mode which propagates simultaneously between the source and the gate and between the source and the drain, and a fast mode less attenuated which propagates mainly between the source and the drain. Although PI topology of our equivalent circuit supports the use of the parameters admittance to characterise our structure, it appears that the impedance parameters are more convenient to calculate.

The expansion of the hyperbolic functions to the order 1 describes an elementary W-wide transistor. Its admittance matrix is $[y].W$

The expansion to the order 3 provides the usual gate and drain access elements

$$R'_g=\frac{r_1 W}{3}\ ;\ L'_g=\frac{l_1 W}{3}$$

$$R'_d=\frac{r_d}{W}+\frac{r_2 W}{3}\ ;\ L'_d=\frac{l_2 W}{3};\ R'_s=\frac{r_s}{W}$$

Lastly the new propagation elements appear with the expansion to the order 5.

$$\begin{cases} Z_{11}=-\dfrac{y'_{22}}{y'_{12}y'_{21}-y'_{22}y'_{11}}\dfrac{1}{W}+\dfrac{z_1}{3}W+\dfrac{z_1^2 y_{11}}{9}W^3 \\[4mm] Z_{22}=-\dfrac{y'_{11}}{y'_{12}y'_{21}-y'_{22}y'_{11}}\dfrac{1}{W}+\dfrac{z_2}{3}W+\dfrac{z_2^2 y_{22}}{9}W^3 \\[4mm] Z_{12}=\dfrac{y'_{12}}{y'_{12}y'_{21}-y'_{22}y'_{11}}\dfrac{1}{W}-\dfrac{z_1 z_2 y_{12}}{36}W^3 \\[4mm] Z_{21}=\dfrac{y'_{21}}{y'_{12}y'_{21}-y'_{22}y'_{11}}\dfrac{1}{W}-\dfrac{z_1 z_2 y_{21}}{36}W^3 \end{cases} \tag{4}$$

That leads to the equivalent circuit in Fig. 4 in which supplementary elements appear, proportional to W3. On the source an impedance $\zeta_{12}W^3$ appears. On the gate an impedance $(\zeta_{11}-\zeta_{12})W^3$ appears. On the drain an impedance $(\zeta_{22}-\zeta_{12})$ W^3 and a voltage source $(\zeta_{21}-\zeta_{12})i_g$ appear. The terms ζ_{ij} are given by:

$$\zeta_{12}=\frac{z_1 z_2 y_{12}}{36}\ ;\ \zeta_{21}=\frac{z_1 z_2 y_{21}}{36}$$

$$\zeta_{11}=\frac{z_1^2 y_{11}}{9}\ ;\ \zeta_{22}=\frac{z_2^2 y_{22}}{9}$$

with $z_1=r_1+jl_1\omega$; $z_2=r_2+jl_2\omega$

We obtain finally the equivalent diagram of a field-effect transistor shown in Figure 4.
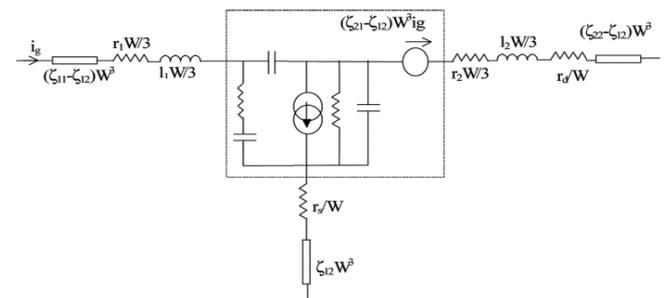
Fig. 4. Equivalent circuit of transistor with the propagation elements

The [Z] parameters of our new equivalent circuit are given in (5) where the matrix [z'] is the inverse of [y'].

$$\begin{cases} Z_{11} = R_g + jL_g\omega + \dfrac{z'_{11}}{W} \\[2mm] Z_{12} = R_s + jL_s\omega + \dfrac{z'_{12}}{W} \\[2mm] Z_{21} = R_s + jL_s\omega + \dfrac{z'_{21}}{W} \\[2mm] Z_{22} = R_d + jL_d\omega + \dfrac{z'_{22}}{W} \end{cases} \qquad (5)$$

with

$$\begin{cases} R_s = \dfrac{r_s}{W} + \Re(\zeta_{12})W^3 \\[2mm] R_g = \dfrac{r_1 W}{3} + \Re(\zeta_{11} - \zeta_{12})W^3 \\[2mm] R_d = \dfrac{r_d}{W} + \dfrac{r_2 W}{3} + \Re(\zeta_{22} - \zeta_{12})W^3 \\[2mm] L_g = \dfrac{l_1 W}{3} + \dfrac{\Im(\zeta_{11} - \zeta_{12})W^3}{\omega} \\[2mm] L_d = \dfrac{l_2 W}{3} + \dfrac{\Im(\zeta_{22} - \zeta_{12})W^3}{\omega} \\[2mm] L_s = \dfrac{\Im(\zeta_{12})}{\omega} \end{cases} \qquad (6)$$

NUMERICAL RESULTS

The following results have been calculated from the founder's data for an OMMIC ED02AH PHEMT.at V_{GS}=-0.1V and V_{DS}=0.2V.

Variation of the Z parameters

Figure 5 shows the variation of W times the real part of Z_{11} as a function of W^2. The slope of this curve for the lowest frequency gives a good approximation of r_1 and its intersection with the ordinate axis gives a good approximation for r_s+r_i. The concavity which appears for larger transistors denotes the effects of the propagation terms.
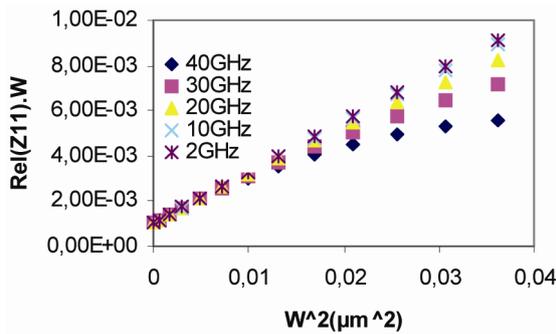
Fig. 5. Variation of $\Re(Z_{11})W$ v.s. W^2 for several frequencies

Figure 6 shows the variation of $W\omega$ time the imaginary part of Z_{11}. The slope is very small, which shows that l_1 is not significant. The intersection with the ordinate axis gives a good approximation for $-1/c_{gs}$.
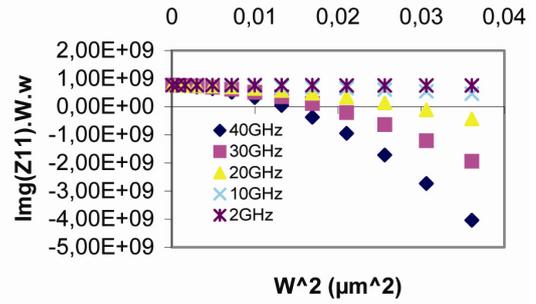
Fig. 6. Variation of $\Im(Z_{11})\omega W$ for several frequencies

The other Z parameters show less significant variations due to the propagation phenomena.

MEASUREMENTS AND RESULTS

The Z-parameters were determined for the OMMIC ED02AH PHEMT with a gate widths (W) were as shown in following.

W (µm)	10	75	100	150	175	200

The above-mentioned transistors were four fingers. It is thus necessary to multiply all parameters Z by four to have only one finger.

Measurements giving the $\Re(Z_{11})W$ and $\Im(Z_{11})\omega W$ variation along the gate width Figure 7 and 8 correspond well to the simulated parameters. However we note that the founder data (Philips Design book) do not correspond with reality.
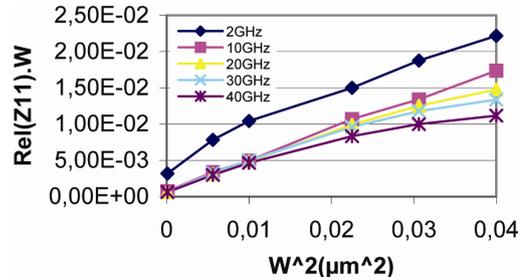
Fig. 7. Variation of $\Re(Z_{11})W$ v.s. W2 for several frequencies
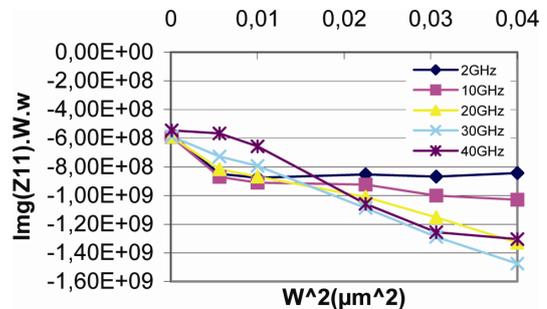
Fig. 8. Variation of $\Im(Z_{11})\omega W$ for several frequencies

The 10μm gate width transistor measurement has sometimes an odd behavior. The same problem for 2GHz frequency.

CONCLUSION

Large field effect transistors are not used because of thermal and propagation phenomena. But, the frequencies becoming higher and higher, the propagation in short transistors may be significant. The results that we present in this paper are limited at 40GHz, which corresponds to the higher frequency of our network analyser. When reaching frequency around 100GHz, we can imagine that the propagation becomes sensitive even for short transistors. This is the reason why we have developed a modified equivalent circuit.

The variations of the Z parameters of a transistor with the gate width allows to evaluate the propagation effects on the equivalent circuit. The concavities of the curves that we have drawn give this information. But some other features of the curves give good approximations for some elements of the conventional equivalent circuit.

A new extraction procedure is being developed and a new noise model is being carried out by the use of the same theory.

REFERENCES

[1] C. Byl, Y. Crosnier. Modeling and optimisation of travelling wave field effect transistors, *Journ. Of Electromagnetic Waves Applications*, (1991), pp. 419-438.

[2] R.L. Chang, T.J. Shieh, W.A. Davis, R.L. Carter. Modeling and analysis of GaAs MESFETs considering the wave propagation effect, *IEEE MTT-Symposium Digest* 1989, pp. 371-374.

[3] W. Heinrich. Limits of FET modelling by lumped elements, *Electronics Letters*, Vol 22 (1986), pp. 630-632.

[4] W. Heinrich, H.L. Hartnagel. Wave propagation on MESFET electrodes and its influence on transistor gain, *IEEE Trans. on MTT,* Vol. MTT-35 (1987), pp. 1-8.

[5] M. Ariaudo, E. Bourdel, D. Pasquet. Variation in intrinsic parameters of transistor with gate width*, Electronics Letters,* Vol. 36, n° 15 (juillet 2000), pp. 1323-1325.

[6] A.J. Holden, D.R. Daniel, L. Davies, C.H. Oxley, H. Rees. Gallium arsenide travelling field effect transistors, *IEEE. Trans. Electrn. Devices*, Vol. 32, No. 1, Jan. 1985.

[7] Y.A. Ren, H.L. Hartnagel. Traveling wave power FET with coplanar slow-wave electrode system modelling and experimental results, *AEU repot.*, 36, pp. 468-474, 1982.

[8] W. Heirich. Distributed equivalent-circuit model for travelling wave FET design, *IEEE. Trans. Microwave theory Tech.,* Vol. MTT-35, No. 5, May 1987.

[9] W. Heirich, H.L. Hartnagel. Field-theoretic analysis of wave propagation on FET electrodes includind losses and small signal amplification, *Int. J. Electron.,* Vol. 58, pp. 613-627, 1985.

PERFORMANCE ANALYSIS OF ASYNCHRONOUS FFH-MA UNDER THE PRESENCE OF THE FREQUENCY OFFSET

Jeungmin Joo, Kanghee Kim, Hyunduk Kang, and Kiseon Kim,
Department of Information and Communications,
Kwangju Institute of Science and Technology, Kwangju, Republic of Korea,
gangsang@kjist.ac.kr

ABSTRACT

In order to investigate the effect of asynchronous fast frequency hopping-multiple access (FFH-MA) systems due to the frequency offset, we evaluate the bit error rate (BER) performance, using noncoherent *M*-ary frequency shift keying (FSK) modulation in the Rayleigh fading channel. While the frequency offset increases at a given signal to noise ratio (SNR), the BER is severely degraded due to the loss of orthogonality of received symbols. With 10% frequency offset, about 5 dB SNR is required additionally to obtain 2×10^{-3} BER, compared to that in the perfectly frequency synchronized case. For the SNR of more than 20 dB, the threshold level of the receiver suffering from the frequency offsets should be greater than that of the perfectly synchronized receiver.

KEYWORDS: *Hysteretic Oscillator, Frequency domain techniques, Floquet's multipliers.*

INTRODUCTION

For several decades, frequency hopping-multiple access (FH-MA) communication systems have become an attractive technique. Both of synchronous and asynchronous FH-MA systems have been used extensively in the areas of satellite and ground military applications. Recently, FH-MA systems have become more common for commercial applications in the license-free industrial, scientific, and medical (ISM) bands [1].

The performance analysis of the various synchronous FH-MA systems have been evaluated in [2]-[5]. In the asynchronous case, however, it is not easy to analyze the system performance due to the difficulties in handling asynchronocity of the multiple interferers. The evaluation of multiple integral to obtain the expression for the probability of error, which is not very attractive, is required. Many literature works related to asynchronous FH-MA systems have been reported. At first, the error probability is upper bounded by 1/2 whenever a hop is hit by interferers in [6]. In [7] and [8], a Gaussian approximation method was used to simply approximate the probability error. In [9], the effect of non-orthogonality of the interfering signals due to the asynchronocity (i.e., random delays) of users was neglected in the analysis on the assumption that the frequency separation between two symbols is much larger than the minimum required to guarantee orthogonality. In [10] and [11], expression for the probability of error was derived when a hop is hit only by one interfering user. Recently, in [12], by using characteristic function and two dimensional integral, the approximated error probability was obtained.

Like any other digital communication systems, the FH-MA system suffers from frequency offsets caused by the imperfect estimate of hopping frequencies in the dehopper. Hence, it is interesting to investigate the effect of frequency offset in asynchronous FH-MA systems.

In this paper, we analyze bit error probability of the fast FH-MA (FFH-MA) system considering all possible hits under the presence of frequency offset.

This paper is organized as follows: Section II describes the system and channel models. In Section III, the BER formulas for the FFH-MA system are derived under the presence of the frequency offset. Numerical results are described in Section IV, and conclusions are given in Section V.

SYSTEM AND CHANNEL MODEL

A. Transmitter Model

The binary data sequence of rate R_b is mapped into an M-ary frequency shift keying (MFSK) signal sequence of rate R_s, where $R_s = 1/T_s = R_b/\log_2 M$. Let the kth user transmit a symbol, m_k, for T_s, where $m_k \in \{0, 1, \cdots, M-1\}$. The symbol representing an MFSK signal is converted to a constant L-tuple representing a $\log_2 M$ bits message, i.e., $\boldsymbol{m}_k = \{m_k, m_k, \cdots, m_k\}$. The message represented as an L-tuple is added modulo-M to an address sequence which is an L-tuple and unique for each user, i.e., $\boldsymbol{a}_k = \{a_{k0}, a_{k1}, \cdots, a_{k(L-1)}\}$, where $a_i \in \{0, 1, \cdots, M-1\}$. As a result, it produces a new sequence of length L, i.e., $\boldsymbol{b}_k = \{b_{k0}, b_{k1}, \cdots, b_{k(L-1)}\} = \boldsymbol{m}_k \oplus \boldsymbol{a}_k$, where $\oplus$ denotes modulo-M addition. Each codeword, b_{kl}, is used for selecting one out of the available M orthogonal frequencies. Each codeword modulated by the MFSK modulator is then hopped to one among Q hopping frequencies per hop duration, $T_h = T_s/L$. Each user is assigned independently a Markov hopping pattern which makes two consecutive hopping frequencies to be different. The signal transmitted by the kth user for T_s is given by

$$S_k(t) = \sum_{l=0}^{L-1} \sqrt{2P} p_{T_h}(t - lT_h) \exp[j(2\pi(f_c + f_{kl})t + \theta_{kl})],$$

(1)

where $p_{T_h}(t)$ is a unit amplitude pulse of the hop duration. P and f_c are the transmitted power and carrier frequency, respectively, and θ_{kl} is random phase introduced by the frequency hopper. $f_{kl} = q_{kl}M/T_h + b_{kl}/T_h$, where $q_{kl} \in \{0, 1, \cdots, Q-1\}$ is a hopping frequency index.

Let K active users exist in the FFH-MA system. According to the frequency hopping pattern and address sequence of the desired user, the received signal is dehopped. When J_l interferers cause hits for the lth hop, the resulting output signal is represented as

$$r_d(t) = \sum_{l=0}^{L-1} \sum_{k=0}^{J_l} \alpha_{kl} \exp(j\phi_{kl}) p_{T_h}(t - (l - \varepsilon_k)T_h)$$
$$\exp[j(2\pi(f_c + g_{kl}/T_h + \Delta f)t + \theta_{kl})] + n(t), \quad (2)$$

where the part corresponding to $k = 0$ is the desired user's signal, and $g_{kl} = (b_{kl} - a_{0l}) \bmod M$. Note that $g_{0l} = m_0$ for all l. Δf denotes the frequency offset. It is assumed that the frequency offset, Δf, is the same for L hops of all users, and less than the minimum frequency spacing among M-ary symbols, i.e., $-1/T_h \leq \Delta f \leq 1/T_h$. α_{kl} is the channel gain having Rayleigh distribution, ϕ_{kl} is the phase of channel having uniform distribution $[0, 2\pi)$, and $n(t)$ denotes the AWGN with two-sided power spectral density of $N_0/2$. Note that uncertainty about phase of the received signal, $\theta_{kl} + \phi_{kl}$, is also uniformly distributed within $[0, 2\pi)$, and independent of α_{kl} [12]. ε_k is timing misalignment relative to the desired signal, where ε_k is uniformly distributed within $[0, 1)$ for $k \neq 0$ and ε_0 is set to be zero.

The dehopped signal passes through the noncoherent MFSK demodulator, which consists of M envelope detectors. The output of the nth envelope detector is represented as

$$R_{nl} = \left| \frac{2}{T_h} \int_{lT_h}^{(l+1)T_h} r_d(t) \exp(j2\pi(f_c + n/T_h)t)dt \right|, \quad (3)$$

where $|\cdot|$ means an absolute value of the argument. We assume that the average received power per hop is the same for all L hops. Then, the conditional probability density function (pdf) of R_{nl} has the Rayleigh distribution given by

$$f_{R_{nl}}(R|\Omega^2) = \frac{R}{\Omega^2} \exp\left(-\frac{R^2}{2\Omega^2} \right), \quad (4)$$

where

$$\Omega^2 = \sum_{k=0}^{J_l} P_{ave} \operatorname{sinc}^2(n_{kl} + \rho) + \sum_{k=1}^{J_l} \Big\{ P_{ave}(1 - \varepsilon_k)^2 \cdot$$
$$\operatorname{sinc}^2((n_{kl} + \rho)(1 - \varepsilon_k)) \Big\} + N_0/T_h, \quad (5)$$

where $\operatorname{sinc}(x) \overset{\text{def}}{=} \sin(\pi x)/\pi x$, P_{ave} is the average received power, ρ is normalized frequency offset, i.e., $\rho = \Delta f/T_h$, and $n_{kl} = g_{kl} - n$. Note that even though interferers' signals are received asynchronously, R_{nl} does not contain interferences caused by adjacent symbols around the lth symbol. The reason is why the Markov hopping pattern is used.

The demodulated signal passes through hard-decision detector (the resulting output is 1 or 0), and the receiver's time-frequency (T-F) matrix can be reconstructed [3, 12].

BER Performance Analysis

By using (4), the insertion probability is obtained, which is used for the hard-decision decoding and majority logic decision. With the majority logic decision, it is investigated that how many entries exist in spurious rows and a desired

row of the T-F matrix. Then, based on the majority logic decision rule, if a desired row contains more entries than any spurious rows, a correct decision is attainable, otherwise, error will occur [3].

A. Insertion Probability

The insertion probability is defined as the probability that the output of the envelope detector is larger in magnitude than a given threshold, β. Let $\boldsymbol{n}_l = \{n_{0l}, n_{1l}, \cdots, n_{J_l l}\}$ and $\boldsymbol{\varepsilon} = \{\varepsilon_0, \varepsilon_1, \cdots, \varepsilon_{J_l}\}$. Conditioned on $\boldsymbol{n}_l$ and $\boldsymbol{\varepsilon}$, the insertion probability at the nth envelope detector for the lth hop can be given by

$$P_{Inl}(\boldsymbol{n}_l, \varepsilon_l) = \int_0^\beta f_{R_{nl}}(R|\Omega^2)dR = \exp\left(\frac{-B}{\chi(\boldsymbol{n}_l, \varepsilon)} \right), \quad (6)$$

where $B = \beta_0^2 N_0/2\bar{E}_h$, and

$$\chi(\boldsymbol{n}_l, \boldsymbol{\varepsilon}) = \sum_{k=1}^{J_l} \chi_k(n_{kl}, \varepsilon_k) = \sum_{k=1}^{J_l} [\operatorname{sinc}^2(n_{0l} + \rho) + N_0/\bar{E}_h]/J$$
$$+ (1 - \varepsilon_k)^2 \operatorname{sinc}^2[(n_{kl} + \rho)(1 - \varepsilon_k)], \quad (7)$$

where $\bar{E}_h/N_0$ is the average received SNR per hop, and $\beta_0 = \beta/\sqrt{N_0/T_h}$ is the actual threshold level β normalized by the rms noise power, which is used commonly in envelope detectors. This threshold level needs to be selected properly to optimize the BER performance. The selection of a threshold level depends on system parameters and channel conditions. (6) contains J_l random variables related to the asynchronous transmission. For simplifying the analysis, the characteristic function can be used. Let $\chi(\boldsymbol{n}_l, \boldsymbol{\varepsilon}) = \chi_l$ and $\chi_k(n_{kl}, \varepsilon_k) = \chi_{kl}$ for simplifying the notation. When $f_{\chi_l}(x|\boldsymbol{n}_l)$ exists over $x_L \leq x \leq x_U$, (6) can be represented as

$$P_{Inl}(\boldsymbol{n}_l) = \int_{x_L}^{x_U} \exp(-B/x)f_{\chi_l}(x|\boldsymbol{n}_l)dx$$
$$= \frac{1}{2\pi} \int_{-\infty}^\infty \Phi_{\chi_l}(\omega|\boldsymbol{n}_l)H(\omega)d\omega, \quad (8)$$

where $\Phi_{\chi_l}(\omega|\boldsymbol{n}_l)$ is the characteristic function of $f_{\chi_l}(x|\boldsymbol{n}_l)$, which is defined as $\int f_{\chi_l}(x|n_l)\exp(j\omega x)dx$, and

$$H(\omega) = \sum_{k=1}^\infty \frac{(-B)^k}{k!} \left\{ \sum_{n=1}^{k-1} \left[\frac{(-j\omega)^{n-1}}{(k-1)\cdots(k-n)} \frac{\exp(-j\omega x)}{x^{k-n}} \right] \right.$$
$$\left. + \frac{(-j\omega)^{k-1}}{(k-1)!} E_i(-j\omega x) \right\} + \frac{j}{\omega} \exp(-j\omega x) \Bigg|_{x=x_L}^{x_U}, \quad (9)$$

where $j \overset{\text{def}}{=} \sqrt{-1}$ and $f(x)|_a^b$ denotes $f(b) - f(a)$. $E_i(\pm jx) = \operatorname{Ci}(x) \pm j\operatorname{Si}(x)$, where $\operatorname{Ci}(x) = -\int_x^\infty \frac{\cos(t)}{t}dt$ and $\operatorname{Si}(x) = -\frac{\pi}{2} + \int_0^x \frac{\sin(t)}{t}dt$ are called Cosine integral and Sine integral, respectively [13].

Since $\{\chi_{kl}|k = 1, 2, \cdots, J_l\}$ are independent random variables, $\Phi_{\chi_l}(\omega|\boldsymbol{n}_l) = \prod_{k=1}^J \Phi_{\chi_{kl}}(\omega|n_{kl})$. Conditioned on n_{kl}, the

pdf of χ_{kl} is given by

$$f_{\chi_{kl}}(x|n_{kl}) = \begin{cases} \frac{[2|n_{kl}|+\text{sgn}(n_{kl})]/|n_{kl}+\rho|}{2\pi\sqrt{\frac{x-a}{[\pi(n_{kl}-\rho)]^2}-(x-a)^2}} & \text{for } a \leq x \leq b \\ \frac{|n_{kl}|/(\pi|n_{kl}+\rho|)}{\sqrt{\frac{x-a}{[\pi(n_{kl}-\rho)]^2}-(x-a)^2}} & \text{for } b < x \leq c, \end{cases}$$

(10)

where $a = [\text{sinc}^2(n_{0l}+\rho) + N_0/\bar{E}_h]/J$, $b = a + \frac{1-\cos(2\pi\rho)}{2(\pi n_{kl})^2}$, and $c = a + \frac{1}{(\pi n_{kl})^2}$. $\text{sgn}(x) = 1$ for $x \geq 0$ and $\text{sgn}(x) = -1$, otherwise. By using Taylor series ($\frac{1}{\sqrt{1+x}} = 1 - \frac{1}{2}x + \frac{3}{8}x^2 - \cdots$) and some manipulations, $\Phi_{\chi_{kl}}(\omega|n_{kl})$ can be given by

$$\Phi_{\chi_{jl}}(\omega|n_{jl}) = \frac{2|n_{jl}| + \text{sgn}(n_{jl})}{2\exp(-j\omega a)} \int_0^{b-a} F(t)e^{j\omega t}dt$$
$$+ |n_{jl}|\exp(j\omega a)\int_{b-a}^{c-a} F(t)e^{j\omega t}dt, \quad (11)$$

where $F(t) = \frac{1}{\sqrt{t-\frac{t^2}{c-a}}} = \sum_{n=0}^{\infty}\left(\prod_{k=0}^{n}\frac{1+2k}{2k+\delta(k)}\right)\frac{t^{-0.5+n}}{(c-a)^n}$ and $\delta(k)$ is the kronecker delta function. (11) can be simplified with the following relationships [13]:

$$\int x^t e^{zx}dx = \frac{x^t e^{zx}}{z} - \frac{t}{z}\int x^{t-1}e^{zx}dx \quad (12)$$

$$\int \frac{e^{zx}}{\sqrt{x}}dx = \sqrt{\frac{\pi}{-z}}\text{erf}(\sqrt{-zx}), \quad (13)$$

where $\text{erf}(x) \stackrel{\text{def}}{=} \frac{2}{\pi}\int_0^x e^{-t^2}dt$. (12) and (13) can be used for evaluating (11) by letting $t = \{-0.5 + n|n = 0, 1, 2, 3, \cdots\}$ and $z = j\omega$. However, it still requires infinite summation to get the exact result of (11) when $F(t)$ is converted to Taylor series. Numerical observation shows that when $F(t)$ is represented as Taylor series, $F(t)$ specified within $0 \leq t \leq (c-a)/2$ can be represented as the summation of at most 10 terms. However, in order to represent $F(t)$ specified within $(c-a)/2 < t \leq (c-a)$, the summation of at least 100 terms is required. In other words, the amount of summation depends on the integral interval, or ρ in (11). To solve this problem, we use the symmetric property of $F(t)$ for $0 < t < (c-a)$. For example, for $t_2 > t_1 \geq (c-a)/2$, $\int_{t_1}^{t_2} e^{j\omega t}F(t)dt = e^{j\omega(c-a)}\int_{c-a-t_2}^{c-a-t_1} e^{-j\omega t}F(t)dt$. When using above relationship, it is enough to use at most 10 terms of $F(t)$ represented as Tayler series.

B. Bit Error Rate

Due to asynchronous transmission and frequency offset, different received MFSK symbols are no longer orthogonal. On the assumption that the mth symbol has been transmitted by the desired user, the insertion probability at each envelope detector is averaged over all transmission symbols of interferers.

$$P_{Inl}(m) = \frac{1}{M^J}\sum_{\substack{\text{all } n_{kl} \\ \text{except } n_{0l}=m-n}} P_{Inl}(\boldsymbol{n}_l) \quad (14)$$

where $\boldsymbol{n}_l = \{n_{0l}, n_{1l}, \cdots, n_{Jl}\}$, $n_{kl} = g_{kl} - n$, and $g_{kl} \in \{0, 1, \cdots, M-1\}$. Then, the insertion probabilities of entries at the desired row and spurious rows are given by $P_{Iml}(m)$ and $P_{Inl}(m)$, respectively, where $n \neq m$.

The probability that there are i entries in the nth row is given by

$$P_n(i|m) = \sum_{\substack{l_1,\cdots,l_i \\ 0 \leq l_1 < \cdots < l_i \leq L-1}} \prod_{x=1}^{i} P_{Inl_x}(m) \prod_{\substack{\text{all } l \in \{0,\cdots,L-1\} \\ l \neq l_1 \neq \cdots \neq l_i}} (1 - P_{Inl}(m))$$

(15)

The probability that i is the maximum number of entries and exactly k spurious rows contain i entries is given by

$$P(i,k|m) = \sum_{\substack{n_1,\cdots,n_k \\ 0 \leq n_1 < \cdots < n_k \leq M-1}} \prod_{x=1}^{k} P_{n_x}(i|m)$$
$$\prod_{\substack{\text{all } n \in \{0,\cdots,M-1\} \\ n \neq n_1 \neq \cdots \neq n_k}} \sum_{k=0}^{i-1} P_n(k|m) \quad (16)$$

Let $\boldsymbol{J} = \{J_0, J_1, \cdots, J_{L-1}\}$ be a given hit pattern for L hops. Conditioned on a given hit pattern, the conditional probability of correct symbol detection is given by

$$P_c(\boldsymbol{J}) = \frac{1}{M}\sum_{m=0}^{M-1}\sum_{i=0}^{L} P_m(i|m)\sum_{k=0}^{M-1}\frac{1}{k+1}P(i,k|m) \quad (17)$$

Then, the probability of correct symbol detection is given by

$$P_c = \sum_{J_0,J_2,\cdots,J_{L-1},} P_c(\boldsymbol{J})\prod_{l=0}^{L-1} P_h(J_l) \quad (18)$$

where the probability that J_l of total $(K-1)$ interferers cause hits for the lth hop is given by $P_h(J_l) = \binom{K-1}{J_l}(\frac{2}{Q})^{J_l}(1-\frac{2}{Q})^{K-J_l-1}$ [13]. Finally, the BER is obtained by $P_b = \frac{M}{2(M-1)}(1-P_c)$

NUMERICAL RESULTS

For the numerical analysis, it is assumed that 4 active users exist in the system, and the separation between adjacent hopping frequencies is assumed to be larger than the coherent bandwidth so that each hopping frequency channel fades independently.

In Fig. 1, the sensitivity of the BER performance in the asynchronous FFH-MA system due to the frequency offset is illustrated using the 4-ary FSK signaling. The frequency offset and asynchronous transmission of users destroy the orthogonality among received symbols, giving rise to significant interference. The increase of the frequency offset induces the increase of loss of the desired signal's energy, so that the BER would be more degraded. For getting the BER of 10^{-2} at given normalized frequency offset of ρ=0.2, the SNR of about 1.5dB is additionally required, compared to the case of the perfect frequency synchronization. With the normalized frequency offset of more than 0.2, however, the BER of 10^{-2} is not obtainable even though the SNR increases. Moreover, with the normalized frequency offset of more than 0.4, it is difficult to achieve the improvement of the BER performance simply by increasing the SNR.

Figure 2 shows normalized threshold levels commonly used in envelope detectors to obtain the BER performance depicted in Fig. 1. A numerical search algorithm is employed for specific values of the SNR and frequency offset to obtain normalized threshold values optimizing the BER performance. The step size for the search algorithm is set to 0.1, because it is small enough to optimize the BER. Such normalized threshold levels depend on given system parameters (i.e., M, Q, and L) and channel conditions. Figure 2 shows that, with the existence of the frequency offset, the normalized threshold level for the large SNR of more than about 20dB is quite high compared to the case of the perfect frequency synchronization (i.e., ρ=0). This implies that when the SNR increases, the amount that the desired user's signal loses its own energy due to the frequency offset becomes larger than the amount that interferers' signals are robbed of their own energy by the desired envelope detector, because most of energy that interferers' signals lose is absorbed into unwanted envelope detectors. Consequently, if the threshold level for the perfect frequency synchronization system is used for the system suffering from the frequency offset, the BER performance of the system with the frequency offset would be more degraded.

CONCLUSIONS

Frequency offsets give rise to the reduction of the desired signal's energy and cause the interference among symbols. The BER performance is severely degraded with the increase of the frequency offsets. For large SNR values of more than 20dB, as the frequency offset increases, the normalized threshold level also increases compared to that used for the perfect frequency synchronization.

ACKNOWLEDGMENT

REFERENCES

[1] S. Glisic, Z. Nikolic, N. Milosevic, and A. Pouttu. Advanced frequency hopping modulation for spread spectrum WLAN, *IEEE J. Select. Areas Commun.*, vol. 18, no. 1, pp. 16-29, Jan. 2000.

[2] G. Cooper and R. Nettleton. A spread pecturm technique for high capacity mobile communication, *IEEE Trans. Veh. Technol.*, vol. VT-27, pp. 264-275, Nov. 1978.

[3] D. Goodman, P. Henry, and V. Prabhu. Frequency-hopped multilevel FSK for mobile radio, *Bell. Syst. Tech. J.*, vol. 59, pp. 259-264, Sept. 1980.

[4] R. Muammar. Degadation in FH-MFSK mobile radio system capacity due to Rayleigh fading and log-normal shadowing, *IEEE Trans. Veh. Technol.*, vol. 37, no. 3, pp. 130-134, Aug. 1988.

[5] P. Yegani and C. McGillem, "FH-MFSK multiple-access communications systems performance in the factory environment. *IEEE Trans. Veh. Technol.*, vol. 42, no. 2, pp. 148-155, May 1993.

[6] E. Geraniotis and J. Gluck. Coded FH/SS communications in the Presence of combined partial band jamming, Rician non-selective fading and multipath User interference, *IEEE J. Select. Areas Commun.*, vol. 5, pp. 194-210, 1987.

[7] E. Geraniotis. Noncoherent hybrid DS/SFH spread-spectrum multiple-access communications, *IEEE Trans. Commun.*, vol. 34, pp. 862-872, Sept. 1986.

[8] J. Wang and M. Moeneclaey. Multiple hops/symbol FFH-SSMA with MFSK modulation and Reed-Solomon coding for indoor radio, *IEEE Trans. Commun.*, vol. 41, pp. 793-801, May 1993.
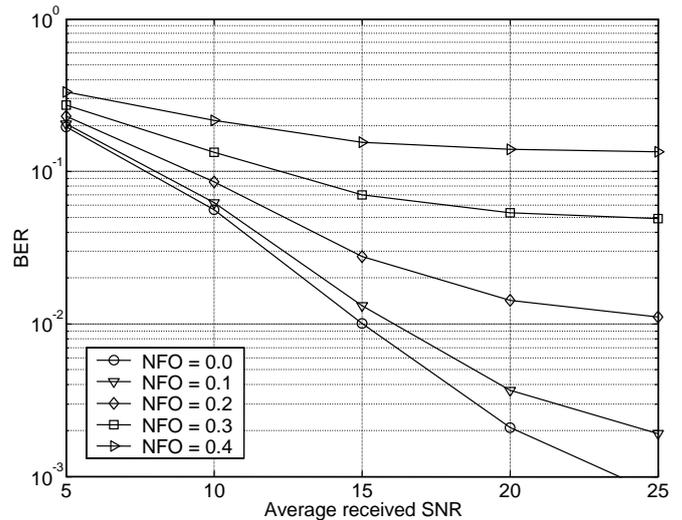
Fig. 1. Sensitivity of the asynchronous FFH-MA system due to frequency offsets for $M = 4$, $Q = 55$, $L = 3$, $N_s/N_c = 1/3$, where NFO means the normalized frequency offset
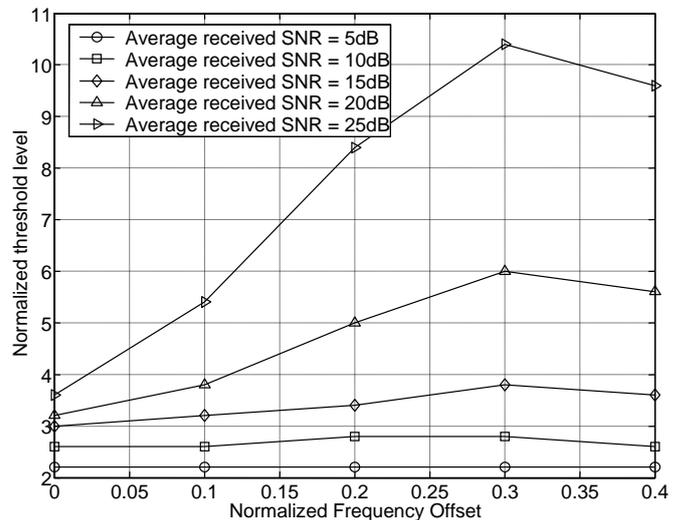
Fig. 2. Normalized threshold levels used in Figure 1

[9] E. Geraniotis. Multiple-access capability of frequency-hopped spread-specturm revisited: an analysis of the effect of unequal power levels, *IEEE Trans. Commun.*, vol. 38, pp. 1066-1077, July 1990.

[10] C. Keller. An exact analysis of hits in frequency-hopped spread-spectrum multiple-access communications, *Proc. Conf. Inform. Sci. Syst.*, pp. 981-986 Mar. 1988.

[11] Q. Zhang and L. Tan. Performance optimization of asynchronous FFH-SSMA systems over multipath Rayleigh channels subject to practical constraints, *Proc. ICCT'96*, vol. 1, pp. 599-602, May 1996.

[12] Y. Su, Y. Shen, and C. Hsiao. On the detection of a class of fast frequency-hopped multiple access signals, *IEEE J. Select. Areas Commun.*, vol. 19, pp. 2151-2164, Nov. 2001.

[13] I. Gradshteyn and I. Ryzhik. *Table of integral, series, and products*, Academic press, 2000.

[14] E. Geraniotis and M. Pursley. Error probabilities for slow frequency-hop spread-spectrum multiple-access communications over fading chan-nels, *IEEE Trans. Commun.*, vol. COM-30, pp. 996-1009, may 1982.

A NOVEL PARALLEL 4X4 TRANSFORM AND INVERSE TRANSFORM ARCHITECTURE FOR H.264

Tiejun Li,
School of Computer Science, National University of Defense Technology, ChangSha, China
tj_li@sohu.com

Sikun Li,
School of Computer Science, National University of Defense Technology, ChangSha, China
lisikun@263.net.cn

ABSTRACT

The H.264 video coding standard provides a compression gain of 1.5 x 2.0 x over H.263 and MPEG-4 simple profile. One of the major differences between H.263 and H.264 is the transform coding. The transforms of H.264 employ only integer arithmetic operations such as additions and shifts without multiplications, with coefficients and scaling factors that allow for 16-bit arithmetic computation on first-level transforms. The integer transforms also solve the mismatch problem like H.263. These changes lead to a significant complexity reduction, but with only less than 0.02 dB decrease in PSNR. A novel parallel 4x4 multiple transforms architecture for H.264 is presented in this paper. This architecture is based on Wallace trees and can process 4 inputs in parallel. This architecture has been designed and synthesized in SMIC 0.18um technology. The result shows that this architecture can achieves above 1,200M pixels/sec throughout and consume only 5625 gates. The timing-area property of this architecture is improved compared with the previous architecture.

KEYWORDS: *Architecture, Transform, H.264, Wallace Tree.*

The article is reworked from unpublished 2nd IEEE International Conference on Circuits and Systems for Communications (ICCSC) materials.

INTRODUCTION

The H.264 video coding standard provides a compression gain of 1.5 x – 2.0 x over H.263 [1] and MPEG-4 simple profile [2]. One of the major differences between H.263 and H.264 is the transform coding [3, 4]. The transforms of H.264 employ only integer arithmetic operations such as additions and shifts without multiplications, with coefficients and scaling factors that allow for 16-bit arithmetic computation on first-level transforms. The integer transforms also solve the mismatch problem like H.263. These changes lead to a significant complexity reduction, but with only less than 0.02 dB decrease in PSNR [4, 5].

Although the computation complexity of the transforms in the H.264 standard is less than H.263, hardware implementation of the transform coding is required in both dedicated and platform codec systems to accelerate the processing and alleviate the loading [6]. Hence, Wang [7] proposed a parallel transform architecture for H.264, which can process 4 pixels per cycle and achieve 360M pixels/sec at 80MHz.

The Wallace tree was proposed by C. S. Wallace in 1964 [8]. This method can be used to sum up all the bits of the partial product in each column and is employed by many fast multipliers [9, 10]. Based on Wallace trees, we propose a new parallel 4x4 transform and inverse transform architecture for H.264 in this paper. The new architecture has the following characteristics:

- It takes Wallace trees instead of adders of four operands and can decrease the length of critical paths and gates count.
- Its four parallel datapaths are more independent and need fewer multiplexers than Wang's.
- The subtracters needed by H.264 transforms are merged in Wallace Trees.

Experimental results show that the area of our architecture decreases about 10% and the critical path delay decreases about 20% compared with Wang's [7].

This paper is organized as followed: Sec.2 introduces the transform coding in H.264. Sec.3 presents the architecture based on adders proposed by Wang[7]. Sec. 4 gives the parallel 4x4 transform and inverse transform architecture based Wallace trees proposed by us. Sec.5 implements and compares the proposed architecture and Wang's.

TRANSFORM CODING IN H.264

The video coding layer of H.264 is similar in spirit to other standards such as H.263. It consists of a hybrid of temporal and spatial prediction, in conjunction with transform coding. Figure 1 shows a block diagram of the video coding layer for a macroblock(MB). The spatial prediction is also called intra prediction. There are two types of intra prediction in H.264. One is 4x4 intra prediction and the other is 16x16 intra prediction. The temporal prediction is the multiple reference frame and variable block size motion estimation. Besides predictions and transforms techniques, H.264 also employ other advanced video encode technique, such as context adaptive variable length coding (CAVLC), Context-based Adaptive Binary Arithmetic Coding (CABAC) and deblocking filter.
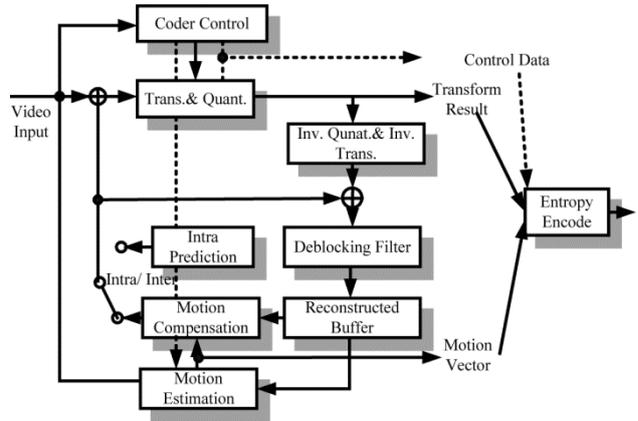
Fig. 1. Block diagram of H.264 encoding flow

The residual MB is coded as Figure 2. It is divided into sixteen 4x4 blocks for luminance and 4 blocks for chrominance. If the 16x16 intra prediction mode is chosen for the MB, the DC items of the luminance are extracted to form a 4x4 block. Hadamard transform is then applied on it.
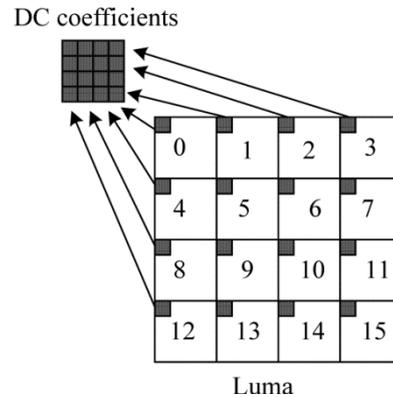
Fig. 2. Residual coding order of H.264

Just like 8x8 DCT, we can implement the 2-D 4x4 integer DCT in this paper by row-column decomposition techniques. There are two types of 4x4 transforms for the residual coding. The first one is for luminance residual blocks. The 1-D transform and inverse transform matrices are shown in (1) and (2).

$$\begin{bmatrix} F0 \\ F1 \\ F2 \\ F3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix} \begin{bmatrix} X0 \\ X1 \\ X2 \\ X3 \end{bmatrix} \quad (1)$$

$$\begin{bmatrix} X'0 \\ X'1 \\ X'2 \\ X'3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1/2 \\ 1 & 1/2 & -1 & -1 \\ 1 & -1/2 & -1 & 1 \\ 1 & -1 & 1 & -1/2 \end{bmatrix} \begin{bmatrix} F'0 \\ F'1 \\ F'2 \\ F'3 \end{bmatrix} \qquad (2)$$

The other type of the transform is Hadamard transform. It is applied to the luminance DC coefficients in 16x16 intra prediction mode. The 1-D inverse Hadamard transform is simply the transpose of (3). Because the transform matrix is symmetric, the inverse hadamard transform is the same as the forward transform.

$$\begin{bmatrix} F0 \\ F1 \\ F2 \\ F3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix} \begin{bmatrix} X0 \\ X1 \\ X2 \\ X3 \end{bmatrix} \qquad (3)$$

The three types of transform matrices above only contain 6 coefficients: 1, -1, 2, -2, 1/2 and -1/2, which can be implemented by shifters and adders. The sign of the coefficient at the same position is same. These characters are helpful to simplify the design of the transform architecture.

TRANSFORM ARCHITECTURE BASED ON ADDERS

Wang [7] implemented a 4x4 integer DCT architecture by row-column decomposition techniques. He overlapped three transforms in (1), (2) in a single architecture, which is showed in Figure 3. In Figure 3, all the adders have three inputs, among which one input is not changed and one of the other two inputs is required to be selected by the transform type. For the different transform types, some inputs should be multiplied by corresponding coefficients: 1, -1, 2, - 2, 1/2 or -1/2.
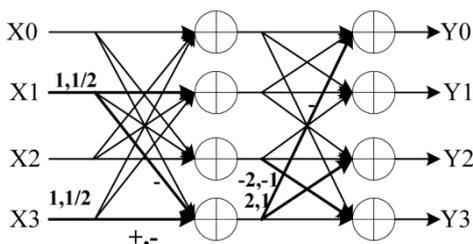
Fig. 3. Architecture for three types transform based adders

PROPOSED ARCHITECTURE

2-D 4x4 Parallel Transform Architecture

We also adopt the classical row-column decomposition method [11, 7] for 2-D 4x4 transform and inverse transform architecture for H.264, as shown in Figure 4.
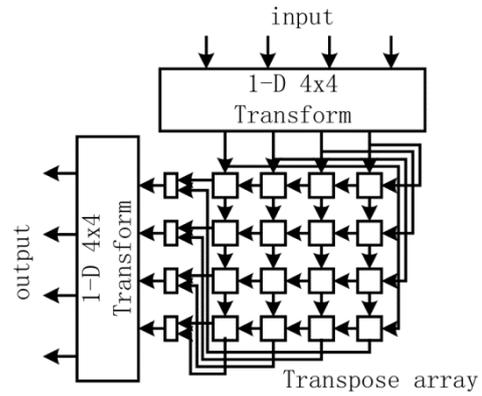
Fig. 4. 4x4 2-D transform architecture

The output of the first 1-D 4x4 transform module on the top of Fig.4 is fed to a transpose array. The transposing result of transpose array is fed to the second 1-D 4x4 transform module as the left transform part of Fig.4. The two transform modules and transpose array form a pipeline structure, which enable the architecture can process 4 pixels per cycle.

The transpose array is a 2-D FIFO array, which can shift along two directions. One FIFO element consists of a two input multiplexer and a register. The multiplexer controls the data flow direction of the FIFO. The first input of the multiplexer is the data from the upper register. The second input of the multiplexer is the data from the right register. If all of the multiplexers select the data from the upper registers, the FIFO array will be shift down. The bottom row of the register array will be outputted to the second 1D transform unit. However, if all of the multiplexers select the data from the right, the FIFO array will be shift left. The direction will be changed every four valid input clocks so that the transpose operation can be done in this register array.

1-D 4x4 Transform Architecture

The 1-D 4x4 transform module is shown in Figure 5, which contains four Wallace trees and eight shift modules.
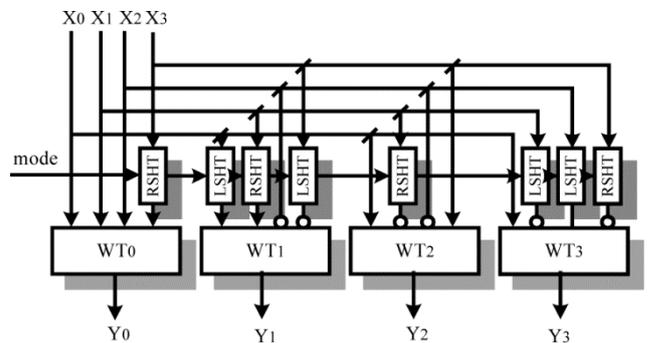
Fig. 5. Architecture for three types transform based Wallace trees

The structure make the best of the characters of 4x4 transforms of H.264 discussed in Sec.2. The four parallel datapaths are independent from each other and intercross little. The core of the datapath is a Wallace tree. Every Wallace tree implements a 4x1 transform. The inputs of the Wallace trees are independent and do not need to be selected by multiplexers. In order to implement the multiplying of 2 and 1/2, two type shift modules are designed: right shifts (RSHT) and left shifts (LSHT). The sign 'o' before some inputs of Wallace trees are designed for subtracting operations.

The shift modules are controlled by the transform mode to decide whether or not to shift and introduce only a delay of one level gate. If the transform mode is the 4x4 transform for luminance residual blocks as (1), LSHTs are selected. If the transform mode is the 4x4 inverse transform for luminance residual blocks as (2), RSHTs are selected. Otherwise, all shifts are all deselected.

Wallace Trees Architecture

The structure of a Wallace tree is shown in Figure 6. In our architecture, Wallace trees are used to substitute for the adder trees of four operands.
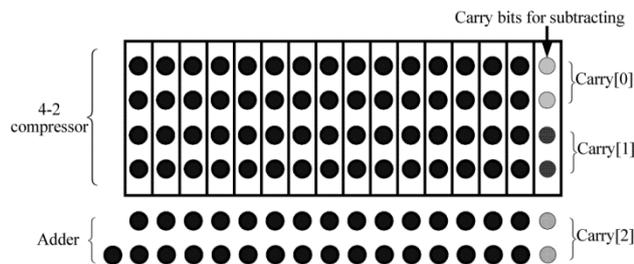
Fig. 6. Wallace tree supporting subtracting

The Wallace trees offer three advantages compared with adder trees. Firstly, a single 4-2 compressor contains only 6 gates and includes a critical path with the maximal delay of three XORs [9, 10], which will decrease the cost of the area and timing. Secondly, the structure of Wallace trees is more regular than adder trees' and is suitable for the placement and routing. Finally, the subtracting needed by H.264 can easily be merged into Wallace trees while introducing a little of gates.

A subtracting can be looked as an addition of the minuend and the negative of the subtrahend. The negative of the subtrahend can be calculated by complementing each bit and then adding 1. In Figure 5, the sign 'o' before the inputs of Wallace trees denotes the operation of complementing. The operations of adding 1 are postponed into Wallace trees.

By observation of Figure 6, we can find that there is one additional bit in the compressor structure and adder, which can generate four types of carrying value (0,1,2 and 3) to add to original calculating of Wallace trees.

This structure consumes a little of gates and introduces the single delay of INVERT (viz. complementing) to implement three possible types of subtracting in 4x4 transforms of H.264.

Bit Width Design

Our proposed architecture should meet the bit width requirements of three types of transform. Both the specification of H.264 [3] and the analysis [7] guarantee that the 16 bits arithmetic is enough, so the datapaths and transpose array in our architecture are implemented with 16 bits.

IMPLEMENTATIONS AND COMPARISION

The proposed architecture in this paper is designed and simulated by Verilog HDL. The logic is synthesized by Synopsys' DC with the Verisilicon standard cell library [12] on SIMIC 0.18um 1P5M process. In order to compare the implementation results to Wang's architecture fairly, the version of Wang's multi-transforms architecture on the same technology and constraint as us is also implemented.

The other individual architectures proposed by Wang are not implement, because the multiple transforms are always required by H.264 but not any individual transform. The gate count and delay are listed in Table 1.

Table 1

Implementation results of our proposed parallel and Wang's

Architecture	Technology	Critical path delay (ns)	Area (gates)
Wang's[7]	TSMC 0.35um	11.35	6538
Wang's	SIMIC 0.18um	5.35	6305
Proposed	SIMIC 0.18um	4.27	5625

CONCLUSIONS

In this paper, we propose a novel parallel 4x4 multiple transforms architecture for H.264. This architecture supports Forward/Inverse transforms for luma residual blocks and DC coefficients used in H.264 and can process 4 inputs in parallel. We have mapped the architecture to SMIC 0.18um technology.

The architecture can run on a system clock of more than 300MHz and achieve 1,200M pixels/sec throughout. Compared with the previous architecture [7], its area decreases about 10% and critical path delay decreases about 20%. The architecture proposed in this paper can be applied to a hardware accelerator or dedicated design for the H.264 video codecvideo codec design of H.264.

REFERENCES

[1] ITU-T Recommendation H.263: Video coding for low bit-communication Version 1November 1995; Version 2 (H.263, January 1998; 3(H.263++), November 2000.

[2] ISO/IEC 14496-2: Coding of audio-visual objects. Part 2: Visual. ISO/IEC JTC1. MPEG-4 Visual version 1, April 1999; Amendment 1 (Version 2), February 2000.

[3] ITU-T Rec. H.264 / ISO/IEC 11496-10 "Advanced Video Coding", Final Committee Draft, Document JVTG050, Pattaya ,March 2003.

[4] Henrique S. Malvar, Antti Hallapuro, Marta Karczewicz, and Louis Kerofsky, Low-Complexity Transform and Quantization in H.264/AVC, *IEEE transactions on circuits and systems for video technology*, Vol. 13, No. 7, July 2003.

[5] A. Hallapuro, M. Karczewicz, H. Malvar. Low complexity transform and quantization, *Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG*, Jan. 2002, Docs. JVT-B038 and JVT-B039.

[6] Shen Cheng-dong, Li Tie-jun, Li Si-kun, The Performance and Complexity Analysis of The H.264 Video Coding Standard, CCVRV'2003, Sep. 2003.

[7] Tu-Chih Wang, Yu-Wen Huang , Hung-chi Fang, and Liang-Gee Chen, "PARALLEL 4x4 2D Transform and inverse transform architecture for MPEG-4 AVC/H. 264", ISCAS 2003, Bangkok, Thailand, May 2003.

[8] C.S. Wallace. A Suggestion for Fast Multiplier, *IEEE, Trans. Electric Computer*, 1964.

[9] Kaihong Sun, Design and Implementation of a Module Generator for Low Power Multipliers, Master's Thesis, Division of Electronics Systems, Depart. of Electrical Engineering, Linköping Univ., Sweden, Sept. 25, 2003.

[10] Gary W. Bewick, Fast multiplication: algorithms and implementation, The Department of Electrical Engineering and the Committee on Graduate Studies of Stanford Univ., Feb. 1994.

[11] A. Madisetri, A.N. Willson. A 100 MHz 2-D 8x8 DCT/IDCT processor for HDTV applications, *IEEE Transaction on Circuits and Systems for Video Technology*, vol.5, no.2, pp. 158-165, Apr., 1995.

[12] VeriSilicon SMIC 0.18μm 1.8V/3.3V I/O Cell Library Databook, Version: 2.0, Aug 1, 2003.

APPLICABILITY OF 5G SUBSCRIBER EQUIPMENT AND GLOBAL NAVIGATION SATELLITE SYSTEMS

Svetlana Dymkova,

Institute of Radio and Information Systems (IRIS), Vienna, Austria

ABSTRACT

5G is the latest generation of mobile cellular technol-ogy, which is designed to significantly increase Internet speed, coverage and reduce the data packet transfer time in wireless networks. 5G brings together all latest and most advanced developments of mankind in terms of communications and IT. This is the limit of existing tech-nologies of microelectronics and data radio transmission. The new generation of 5G mobile communication has a number of fundamental advantages compared to 4G: higher data transfer rate; low signal delay; the abili-ty to connect more devices; high energy efficiency; multiply increased throughput; high user mobility. Another im-portant difference of 5G deserves attention — large-scale virtualization. The new technology goes beyond just hardware solutions. Many functions in it are imple-mented not at the level of physical infrastructure, but in a soft-ware way. This article covers the topic of applicability of GNSS and 5G mobile communications. The material of article will tell you why GNSS and 5G are interesting and how User Equipment developers can start shaping the 5G device market today.

KEYWORDS: *GNSS, 5G network, standards, MIMO, LTE.*

INTRODUCTION

Communication networks of the fifth generation (5G), together with Big Data and the Internet of things (IoT), are designed to become one of the digital economy foundations, the main driving force of which is artificial intelligence. Over the past 40 years, four generations of mobile networks have changed. If the first generation 1G cellular networks have long disappeared, then 2G, 3G and 4G networks still continue to be exploited. Moreover, a certain amount of the legacy infrastructure of 3G and 4G networks will organically become part of the fifth generation 5G mobile networks [5, 19].

Number of devices connected to the Internet and requirements of subscribers for the speed of mobile access are increasing every year. Developers of telecommunications equipment and telecom operators update the networks architecture and interaction regulations. 5G networks are emerging, which are the next stage in the 4G networks evolution. Geography of 5G application in the world is noticeably expanding. In 2021, we managed to find 5G practices in 14 new countries: France, Italy, Australia and New Zealand, the countries of South America. The largest number of new practices were found in China (22), Great Britain (17) and the USA (16), which became more visible on the global landscape largely due to the activity of local operators [8].

The internal geography of 5G is expanding - they are being actively implemented not only in capitals and megacities, but also in provinces, regions and small towns.

The key functional areas in which 5G practices are being most actively implemented are: industry, transport, entertainment, security, communications and medicine. In industry, there is a development of scenarios that were discovered last year. This is implementation at industrial facilities (at the level of infrastructure hubs), equipment management. In the transport segment, among other things, technically complex practices were implemented: for example, providing communications for high-speed transport routes. In most areas, another trend is noticeable – many new practices have become more narrow-profile, that is, within the framework of one practice, a narrower range of tasks is solved and a smaller range of technologies is used in addition to 5G.

Enterprises are actively using private 5G networks. As a rule, they are deployed by industrial enterprises - most projects in this direction were carried out in private networks. However, such practices are also found in the segments of transport, medicine, security, and agriculture [8]. Half of Russian practices are also implemented in private 5G and 5G-ready networks (that is, ready to be launched in fifth-generation networks after the approval of the relevant standards), also mainly in industrial enterprises [12-14, 16-17].

5G wireless communication networks require very tight time synchronization at every node to maximize spectrum usage efficiency and capacity throughput [18]. All Remote Radio Heads (RRH) must be time synchronized to +/- 1.5 us in their base mode of operation. If your RRH timing drifts outside of this range, it will shut down to avoid interfering with adjacent cells.

5G communications networks rely on high-precision source clocks that lock to the Global Navigation Satellite System (GNSS). If the GNSS timing reference is lost or impaired, calls will drop and data services will dramatically slow down or stop altogether. Highly dense 5G architectures often require a 10× increase in the number of nodes and GNSS traceable source clocks, but ideally, operators could avoid the expense and security exposure of installing GNSS receivers in every clock node [1].

This article covers the topic of applicability of GNSS and fifth generation (5G) mobile communications. The material of the article will tell you why GNSS and 5G are interesting how User Equipment (UE) developers can start shaping the 5G device market today [6].

GNSS & RTK

Let's look at the history of Global Navigation Satellite System (GNSS) and Real Time Kinematic (RTK). One technology that GNSS often relies on, RTK, is a global satellite positioning technique that helps GNSS improve the validity and accuracy of target data. In terms of positioning, positioning and maximum accuracy, the combination of GNSS with RTK increases the level of accuracy. RTK amplifies the phase signal exchanged between transmitter and receiver, providing centimeter-level accuracy and real-time signal correction.

Currently, GPS, GLONASS and Galileo are the main operating GNSS. With the advent of GNSS technology, many assistive technologies known as Regional Navigation Systems came into play. The technology concept is the same as GNSS, but covers less geographic areas.

GNSS satellites have two carrier waves fixed in the L band, namely L1 (1575.42 MHz) and L2 (1227.60 MHz). The main purpose of these two wavebands is to transmit the signals of the connected satellite to the surface of the earth. Using L-band technology can reduce overhead while providing a reliable connection that is less prone to interruptions. The introduction of L-band with the right antenna placement offers a number of benefits for agricultural drones, marine technology, remote monitoring, and more. On the other hand, GNSS receivers placed on the earth's surface consist of an antenna and a processing unit. The purpose of the antenna is to receive coded signals from connected satellites, and the task of the processing unit is to decode the signals into meaningful information.
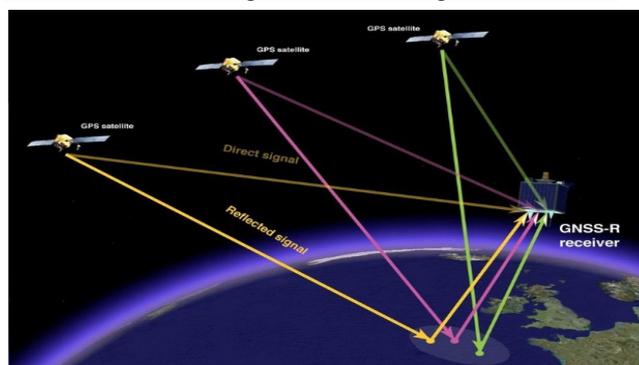
To determine the position of one receiver, GNSS must collect data from at least three separate satellites. Each GNSS satellite revolves around the Earth at an interval of 11 hours 58 minutes and 2 seconds. The time information transmitted by the satellite is transmitted using codes so that the receiver can determine the time interval during which the code was transmitted.

The signals transmitted from a satellite contain coded data that helps receivers pinpoint its location, and the receiver itself positions itself exactly according to the position of the satellite. The receiver IC calculates the time difference between the broadcast time and the coded signal reception time. Once the receiver is positioned accurately relative to the satellite, the processing unit translates the receiver's location in terms of latitude, longitude, and altitude.

The application of GNSS technology has led to a change in the concept of position tracking with a high degree of accuracy and a wide range of coverage. Recently, GNSS has been included in the automotive industry, now almost every car company is integrating GNSS technology into their car models. GNSS is also widely used in aircraft. Preliminary terrain mapping and real-time GNSS terrain updates allow pilots to avoid collisions in air traffic. What's more, the GNSS used in aircraft cockpits also uses technologies such as WAAS or GBAS (LAAS) to improve heading accuracy.

Unlike traditional ground-based navigation aids, the Wide Area Augmentation System (WAAS) provides navigation services across the entire National Airspace System (NAS), according to the Federal Aviation Administration. WAAS provides additional information to GPS/WAAS receivers to improve the accuracy and integrity of current position estimates.

In addition to the wide range of GNSS applications in automobiles and aircraft, GNSS is also used to navigate boats/yachts and ships on the surface of the water. Ships also use a GNSS functional block called the "Man Overboard" (MOB). This feature allows the ship's crew to accurately mark the location of a person who has fallen overboard.

Surveying and geological mapping is another important application of GNSS. Most GNSS receivers use signal data generated at the L1 wave frequency to perform geological mapping. It is equipped with an accurate crystal oscillator that helps the wave reduce clock errors when mapping. Researchers can also make highly accurate measurements by calculating the appropriate offset between GNSS sensors. For example, if an actively deforming area (volcano) is surrounded by multiple receiving stations, then GNSS can be useful to detect any kind of earth deformation or movement.

In addition to the above GNSS applications, important ones also include: mobile satellite communications, location-based emergency and accurate services, improved weather forecasting, photographic geocoding, marketing, etc.

The Inertial Measurement Unit (IMU) plays a vital role in global navigation satellite systems. If the signal is obstructed by something (trees, buildings, etc.), then the signal cannot provide accurate positioning. In tandem, GNSS and IMU provide end users with more powerful and accurate navigation solutions.

Thanks to recent technological advances, many concepts and technologies have significantly changed the playing field for robotics, satellite communications and navigation as we know them. The Global Navigation Satellite System is a key player among innovative technologies. Real Time Kinematic provides GNSS with centimeter-level accuracy with real-time signal correction capability. The combined use of GNSS and RTK ensures maximum accuracy and the highest tracking quality, making it the most powerful combination on the market today.

APPLICABILITY OF 5G SUBSCRIBERS

5G is not just a new standard for mobile communication, the introduction of 5G networks will transform our perception of the world in the long term and lead to the social transformation of society. At the same time, the economics of networks will change: the average transmission speed will increase by 40 times, while the cost of delivery, on the contrary, will decrease by 30 times. By 2024, according to analysts, up to 30% of mobile traffic will go through devices with 5G support. 5G technology will account for 15% of the global mobile telephony sector by 2025 (GSMA forecast, www.gsma.com); for Europe and China this figure will be 30%, and for the USA 50%.

5G standardization

The standardization of 5G technologies and solutions should be completed by 2021, so the term 5G so far refers to only fragmented solutions that will be part of the full-scale IMT2020 solution in the future. Such solutions are already being deployed in different countries, but they are still local and test in nature and do not provide all the planned functionality of IMT2020 networks.

Main standards organizations 5G

3GPP (3rd Generation Partnership Project) – an alliance of seven organizations that develop various telecommunications standards, which, in turn, include other partners. The task of 3GPP is to formulate specifications, evaluate proposals, and finally adopt standards. In addition to developing a common architecture, 3GPP is also developing 5G New Radio (NR) radio technology standards for new frequency bands dedicated to 5G.

ETSI (European Telecommunication Standard Institute), The European Telecommunications Standards Institute, which is a member of 3GPP and is most active in the development of 5G standards.

IETF (Internet Engineering Task Force) develops IP protocol upgrade solutions to support Network Fucntion Vitralization (NFV) network functions virtualization. For example, the IETF has developed Service Function Chaining (SFC) technology that combines virtualized 5G architecture components such as base stations, service gateways, and data packets in a single path. This allows the dynamic creation and chaining of Virtual Network

Functions (VNFs). The IETF works closely with the 3GPP.

ITU (International Telecommunication Union) – a UN agency based in Geneva that standardizes a wide range of telecommunications technologies. In particular, it coordinates work on sharing the radio spectrum, including for 5G networks. In addition to these three main coordinating organizations, there are a number of others in which systematic practical work is underway to develop IMT2020 (5G) standards.

5GPPP (5G Infrastructure Public Private Partnership), is considered one of the leading 5G standardization partnerships. The organization sets ambitious goals for developing requirements for the 5G network, for example, capturing network capacity by 1000 times, reducing the power consumption of user devices by 90%, significantly reducing the time to create new services and services, complete and secure network coverage and with negligible data transfer latency, etc.

NGMN (Next Generation Mobile Networks) Alliance. The Next Generation Mobile Network Alliance is standardizing the full range of 5G solutions. The alliance includes the management of leading American operators: AT&T, U.S. Cellular and Verizon.

In addition to these, there are industry and regional organizations such as 5G Americas, Small Cell Forum, which also make a great contribution to the development and standardization of 5G solutions (Fig. 1, 2).

Figure 1. 5G standardization roadmap in 3GPP

Source: The ABC's of 5G New Radio Standards.
5992-3406EN © Keysight Technologies [10]

Figure 2. Detailed timeline & Process for IMt-2020 in ITU-R

Source: www.itu.int [11]

Large telecom operators such as AT&T, Verizon, etc. also make a great contribution to the development of standards. They coordinate their work with ETSI and ITU, but sometimes they are ahead of these organizations. Therefore, the decisions of these operators often form the basis of ETSI and ITU standards.

The figure below shows the planned distribution of 5G low-frequency spectrum in various countries and regions of the world according to WRC-15 (Fig. 3).

Figure 3. 5G low-frequency band distribution in various countries and regions of the world

Source: Nokia [9]

Key indicators of the 5G standard and technology

5G (fifth generation) – a generation of mobile communications operating in accordance with telecommunications standards following the existing LTE (4G) technology.

The 3GPP consortium began the formation of the 5G-NR (NR – New Radio, radio access technology for 5th generation mobile networks) specification in 2015. Then plans for the preparation of specifications were announced. In accordance with these plans, the 1st phase of the specifications was to be completed before the second half of 2018 (as part of 3GPP release 15), and the 2nd phase – before December 2019 (as part of 3GPP release 16). At the moment, the 1st phase is completed with a delay of a year, and the 2nd is shifted to the third quarter of 2020 (Fig. 4).

Figure 4. 3GPP Specification Release Plan

Source: https://www.3gpp.org

- The 3GPP standards and specifications are created by market participants and take into account a wide variety of business problems, each of which, of course, has its own specific requirements. Recommendation 3GPP TR 38.913 defined the following key indicators for next generation networks:
- 20 Gbps downlink peak data rate (spectral efficiency 30 bps/Hz);
- 10 Gb/s peak uplink data rate (spectral efficiency 15 bit/s/Hz);
- minimum delay in the radio access subsystem for URLLC services – 0.5 ms, for eMBB services – 4 ms;
- maximum density of devices from the IoT world connected to the network in urban conditions – 1'000'000 devices/km^2;
- autonomous operation of devices from the world of IoT without recharging the battery for 10 years;
- support for mobility at a maximum speed of 500 km/h.

And now briefly about some of the technologies through which the actual implementation of fifth generation networks becomes possible.

FREQUENCY AND BANDWIDTH

The 3GPP TS 38.211 V1.2.0 (2017-11) specification defined new radio frequency bands for 5G (Tabl. 1) and divided them into two frequency blocks: FR1 (frequencies up to 6GHz or sub6G) and FR2 (frequencies above 6GHz or mmWave). Working on higher frequency ranges allows you to eliminate various interferences in the network that distort data transmission. In addition, the higher the frequency, the higher the bandwidth, and the bandwidth of the channel directly depends on it. So, for the FR1 block, depending on the SCS used (Sub-Carrier Spacing, subcarrier radio frequency spacing option), the width of one radio channel is up to 100 MHz, for the FR2 block – from 50 to 400 MHz! Unlike LTE networks, which allow channels with a width of only 1.4, 3, 5, 10, 15 and 20 MHz. And if you combine channel width with frequency aggregation (CA), then for one connection you can achieve spectrum of 2 GHz and more (Tabl. 2).

Table 1

Frequency bands for 5G networks

RF Block	RF range
FR1	450 – 6 000 MHz
FR2	24'250 – 52'600 MHz

MASSIVE MIMO AND BEAM FORMING

Beamforming with MIMO antennas is not a new concept and already exists in the cellular market as AAS (Active Antenna System) [15, 20]. An AAS MIMO antenna mounted on a tower allows you to divide the coverage area into static cells, thereby increasing the efficiency of spectrum use, and therefore increasing the number of channels. But today's congested networks require dynamic digital beamforming to maximize spectrum efficiency.

The application of the concept of MIMO antennas in the FR2 millimeter wave becomes even more interesting because millimeter-wave radio waves have good directivity due to the increase in the number of antenna elements per antenna. An array of such antenna elements (256 or more) can be combined into one so-called. Massive MIMO antenna. By controlling the phase and amplitude of the signals, such an antenna is able to dynamically form many strong and sharp beams in the directions of specific users [25, 30]. With Massive MIMO we get:
− strong signal at the output in the direction of the UE;
− s rong signal/noise level in the direction from the UE;
− absence of intercellular interference;
− a signific nt increase in the number of communication channels per cell.

Thus, MIMO technology takes on different meanings in the sub6G and mmWave bands, as shown in Table 2.

Table 2

MIMO in sub6G and mmWave bands

	Sub6G	mmWave
MIMO order	up to 8x8	x2
Meaning	Static spatial multiplexing for multiple users	Dynamic beamforming for one user
Charac-teristic	Multipath, ideal for spatial multiplexing. Extended coverage area, indoor coverage	Distribution in line of sight. Bulk connections with ultra-wide bandwidth

Sounding Reference Signal (SRS)

Known since 3GPP Release 14, the technology is an important addition to Beamforming. It allows the base station to learn about the channel quality through a special packet sent from the UE. Typically, most UEs can only support sending SRS through their main transmit antenna. Therefore, the base station can only receive channel information for this antenna. However, using the transmit antenna selection technique, complete channel information of all UE antennas can be obtained. Therefore, the base station can generate a beam in the direction of the UE in the best way. As a result, the throughput of the UE will increase significantly, especially at points at long and medium distances from the base station (up to +40%).

SCENARIOS AND EXAMPLES OF THE PROVISION OF MOBILE SERVICES IN 5G NETWORKS

Some of the previously listed indicators, such as, for example, peak data transfer rate and autonomy, are simply incompatible and even mutually exclusive. But all these indicators at once should not be performed by one device at a time or, in principle, supported by the entire list. The idea is to distinguish between different types of mobile radio service scenario depending on the degree of importance (high, medium, low) of a particular indicator. In the concept of Network Slicing, the physical architecture of 5G will be divided into many virtual networks or layers, each of which is designed for a different use case.

Each of the scenarios will satisfy one or another set of previously indicated indicators and, accordingly, is aimed at its own market segment. The specification defines only three scenarios:

– e BB (enhanced Mobile Broadband), ultra-broadband mobile communications;

– U LLC (Ultra-Reliable Low Latency Communication), ultra-reliable, low-latency communications;

– MTC (Massive Machine-Type Communications), mass machine-to-machine communication.

NB-IoT and eMTC for mMTC

mMTC – scenario of machine-to-machine interaction, when human participation is minimal, and all processes are automated. mMTC devices include: water, gas, electricity meters; street lighting controllers; parking space sensors; GPS/GLONASS bookmarks; various smoke/fire sensors; burglary sensors; "smart" trash cans and other IoT devices. As you can see, high speed and ultra-low latency are not at all important here, but autonomy and a huge number of connections in the network are very important. We are talking about the so-called. LPWA (Low Power Wide Area) devices – about mass, simple and cheap devices with ultra-low consumption, capable of operating on a single battery up to 10 years.

Standards and specifications for LPWA networks were laid down in releases 13 (Cat.NB1 and Cat.M1) and 14 (Cat.NB2 and Cat.M2) of 3GPP and currently NB-IoT networks (aka LTE Cat.NB1/NB2) and eMTC (LTE Cat.M1/M2) are already in commercial operation. Networks for such devices are characterized by low transmission rates (up to 150 kbps in LTE Cat.NB2 and up to 1 Mbps in LTE Cat.M1), wide and "deep" coverage. It should be noted that the beauty of NB-IoT and eMTC is that the deployment of networks by mobile operators does not require huge investments and the allocation of separate frequency bands – these LPWA networks can operate in existing frequency bands and on existing network equipment, while one basic the station can serve a wider area than existing 2G, 3G or LTE networks.

Streaming video

Market participants predict a shift in such a classic application as "video streaming" to the right, towards increasing data transfer rates without special requirements for delays. The main driver for this will be the need for high-quality 8K video.

Today, there are TVs with 4K video support on the market and some providers provide video content of this quality. But reliable access to such content can only be obtained by those connected to the fiber-optic Internet, access to which is not available in all settlements. With the advent of 5G, 4K and even 8K video will become the norm for all residents of the city and suburbs, and in the field of film/photo production, such quality as detail will become even more important (Fig. 5).
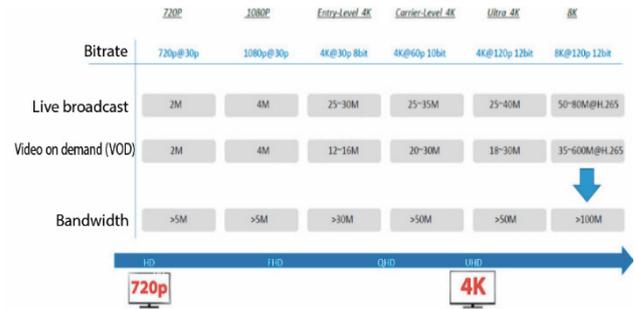

Figure 5. Network bandwidth requirements of different video formats

Consumption of video content on a widescreen TV sets download bandwidth requirements. However, 5G opens up higher speeds for uploads as well. This will open the door to the introduction of urban video surveillance systems with intelligent face recognition on all continents. In such systems, the entire computing part with artificial intelligence is on the network, all that is required of CCTV cameras is to be able to transmit video of the proper resolution to the server. There are examples of implementation of such systems in the world.

The government of Shanghai (China) has been using such a system since 2015. More than 170 million "smart" video cameras are connected to it. For example [2], this system helped to detect a criminal in a crowd of 50,000 on the way from a concert of a popular singer. He came to the concert with his wife and, according to the detainee, he expected to get lost in the crowd.

In practice, such systems bring the city not only savings on security and operational-search activities, but also generate a positive socio-economic effect – citizens and tourists are not afraid to buy expensive things, visit public places at any time of the day, and business is not afraid of the safety of clients and property is now the task of the city. With the advent of 5G, this system has only become more efficient and less costly to deploy and maintain, and therefore more affordable.

Sky Office

In the early stages of commercial 5G deployment, with the exception of smartphones, it is expected that the laptop with Sky Office connectivity will be the key 5G product. Sky Office is the concept of moving laptop computing power to the cloud by equipping the laptop with a built-in 5G modem. The cloud can host not only user files (Cloud Drive), but also software such as MS Office 365 (Cloud Office) or gaming software products (Cloud Games). In this concept, the laptop becomes, simply put, a screen with a keyboard and a camera.

If cellular networks provide a delay of a few milliseconds and provide a dedicated reliable communication channel on an unlimited basis (Network Slice), then working with Sky Office in the future may become a popular way to use a laptop.

At the same time, the consumer will receive a number of interesting consumer qualities that are unattainable with conventional laptops:

• low consumption at the level of tablets with a battery life of 14 hours or more;

• "always ready", the laptop does not waste time downloading software, it is already running - in the cloud;

• "ready everywhere", losing a laptop no longer means losing data and licenses;

• thin and light body, the composition and structure of the notebook are simplified, and this entails a reduction in size and weight;

• passive cooling, the laptop no longer performs energy-intensive calculations and heats up slightly;

• communication is safer than Wi-Fi, because. 5G is almost impossible to hack, the communication channel is protected by the latest encryption algorithms.

Of course, turning the Sky Office concept into reality requires building an entire ecosystem with the participation of players from several industries at once, such as operating system and software manufacturers, laptop manufacturers, mobile operators, cloud service providers, chipset manufacturers, eSIM manufacturers and 5G modules. But despite the complexity of implementation, Sky Office is expected to have rapid growth in China and many other countries in the near future.

Virtual and augmented reality

The entertainment industry has always been a driving force in the development of consumer electronics. The highest performance requirements come from game console consumers. The most advanced, but also less common technologies in the gaming world are virtual reality (VR) and augmented reality (AR).

The well-known companies Sony and Microsoft have been offering accessories for VR and related 3D games for several years now.

Gradually, VR and AR will go beyond the gaming industry and will inevitably spread to education, medicine, industry – it is difficult to overestimate the potential. Figures 10-13 show some examples of the use of AR from the Microsoft Hololens 2 presentation materials. The next step in this industry will be the combination of AR and VR with 5G. Technically, this is already feasible thanks to the new Qualcomm Snapdragon XR2 chipset, which combines a 5G modem and a specialized XR (from VR + AR) processor with artificial intelligence support that responds to the facial expressions of the "pilot".

It is clear that online games will only be acquired with 5G. With the transfer of computing power to the cloud (Cloud Gaming), game consoles will become less loaded, this video will become smoother, more detailed and more dynamic. Having overcome the technological barrier with 5G, the AR/VR games market will become more in demand. Many will discover virtual travel to other cities, diving to the bottom of the ocean and even flying into space. It is a well-known fact that the perception of the world by a person strongly depends on what he sees, with XR + 5G the horizons of the average layman will significantly expand, change the approaches of society to the study of the world and creative activity in all areas.

Tactile internet

In continuation of the topic of XR and artificial intelligence, it must be said separately about the derivative direction, like the tactile Internet. Tactile Internet (hereinafter referred to as TI) is the transmission of tactile sensations, touches over any distance with a minimum, almost imperceptible delay. The name of the technology was proposed at the Dresden University of Technology, where back in 2012, work began on the creation of robotic systems capable of remotely transmitting sensations.

Now scientists are working on creating artificial touch by inserting sensors into soft robotic structures and the most sensitive touch sensors. Now sensors are already able to reproduce the force and nature of the touch, they distinguish between different materials: metal, wood, textiles, etc.

TI puts forward requirements that 5G networks will be able to do:

• delays less than 1 ms;

• reliability – to perform critical tasks (remote operation), network losses, equipment failure, etc. are unacceptable.;

• high data transfer rate – more than 10 Gbps;

• high network density – support for connecting more than 100 devices per 1 km^2.

To reproduce sensations, it is assumed that there are any devices that are in contact with the recipient, for example, clothing (t-shirts, jumpers, trousers), accessories (gloves), shoes, hats, exoskeletons or special devices, which are tactile displays with tiny actuators, which set moving elements (needles, pins) in motion.

With the help of TI, you can learn to draw, play musical instruments, perform remote surgical operations, i.e., everything that requires "fine motor skills". In e-commerce, you can use this technology to touch or try on a product before buying it. You can touch museum exhibits and even feel the weight of ancient artifacts in your hand. Multiplayer online shooters with XR + TI will become more realistic, it will be possible to feel pain, pushes, blows, heat and cold.

The first practical examples of the use of TI in surgery already exist today. In the USA, tests are being carried out for the introduction of the so-called. "telesurgery", when a surgeon performs a surgical operation remotely, via a 5G network. Telesurgery is very different from classical telemedicine – it's not about a simple video broadcast in conference mode, but about the "presence" of the surgeon during the operation. His movements, precision, personal skills, instant reaction to events - everything will be transmitted over 5G networks without a physical presence and without degrading the quality of the operation. Thus, the services of rare specialists will become more accessi-

ble, and patients will be able to choose a surgeon regardless of their country of residence.

Unmanned aircraft (drones)

Telesurgery places high demands on latency and communication reliability, but there is another area that requires, among other things, mass connectivity – unmanned aerial vehicles or "drones". Today, you will not surprise anyone with light unmanned drones for a variety of purposes – from entertainment to specialized military drones. With their help, they shoot spectacular videos, conduct reconnaissance, save people, transport goods, etc. But almost all of them are controlled directly by a person who has a direct wireless reliable contact on an unlicensed frequency.

In the context of the introduction of 5G, in progressive countries, regulatory authorities have already paid serious attention to this topic today, in connection with which work is underway to standardize and ensure security in this area. For example, in Europe there is a special expert group 5G PPP (5G Infrastructure Public Private Partnership, www.5g-ppp.eu/5gdrones) based on the European Commission and representatives of the information and communication technology industry (operators, providers, institutions, small and medium business) from Britain, France, Switzerland, Austria, Finland, Greece, Poland and Estonia. The 5G PPP public-private partnership will offer solutions, architectures, technologies and standards for drones. Through this state initiative, the European Union sees one of the ways to strengthen its technological leadership on the world stage.

With standards to regulate the mass circulation of drones, artificial intelligence systems, a reliable, constant and fast 5G wireless link for a whole drone hive, new markets and services can be opened in a variety of areas. Imagine: drone couriers delivering food from stores or important medicines to hard-to-reach places; rescue drones looking for people lost in the forest or the sea day and night; drones-firefighters, extinguishing fires at an early stage; agrocopters spraying crops – and all on a global scale, and not in particular cases.

Infrastructure C-V2X

Let's move from UAVs to unmanned vehicles. Many have seen the video presentation of Tesla (www.tesla.com), where an electric vehicle controlled by artificial intelligence moves around the city with minimal driver participation. Or another example is the Waymo service (www.waymo.com), which allows you to call a taxi using a mobile application and drive it to a selected point without a driver behind the wheel.

Both services are built on different principles of operation, under the control of powerful artificial intelligence built into the car. Auto makes a situational decision based on visual information and data from lidar (Waymo). A "smart" car is surrounded by "not smart", unpredictable cars controlled by a person.

There is an infrastructure approach to autonomous driving, enshrined in the 14th release of 3GPP – C-V2X. The abbreviation C-V2X stands for Cellular Vehicle-to-Everything, which is the concept of transferring information from a vehicle to any object that can affect the vehicle, and vice versa. This approach allows the vehicle to "communicate" with other cars (V2V), infrastructure (V2I), LTE network (V2N), power grid (V2G), pedestrians (V2P) and even houses (V2H). The 15th release of 3GPP also introduced the possibility of auto communication with the 5G network, which made the C-V2X more attractive thanks to the URLLC service.

Thus, vehicles connected to the C-V2X system will be able to "see" the whole picture of the road situation, "know" about the relative position, obstacles, dangerous areas, and the artificial intelligence located in the network will not just form a trajectory for them separately, but will do this taking into account the mutual influence on the transport system. Such systems will solve the problem of transportation better and safer than any driver, reduce travel time for each participant in the movement, make traffic predictable, safe and energy efficient.

The international consulting company Pricewaterhouse-Coopers (PwC) predicts that the first driverless cars will appear on public roads as early as 2021, and by 2040 all transport in megacities around the world will become unmanned. However, at first, such vehicles will require attention from the driver in certain situations along the way. During this period, legal issues related to unmanned and electric vehicles will also be resolved. In particular, legal and insurance aspects. A certain amount of time will be spent on creating a network of charging stations for electric cars.

Now, having listed many examples where 5G will come in handy more than ever, let's figure out what the state of 5G networks is today and what barriers need to be overcome on the way to a fantastic future.

State of 5G networks in the world

The process of introducing 5G networks into commercial operation has already begun in 2019, however, so far the coverage of such networks is very modest (Fig. 6). At the beginning of 2020, 5G networks were put into operation by 47 operators in 22 countries of the world, and together with those who planned to launch or are testing, there will be 279 operators in 109 countries (Fig. 7).

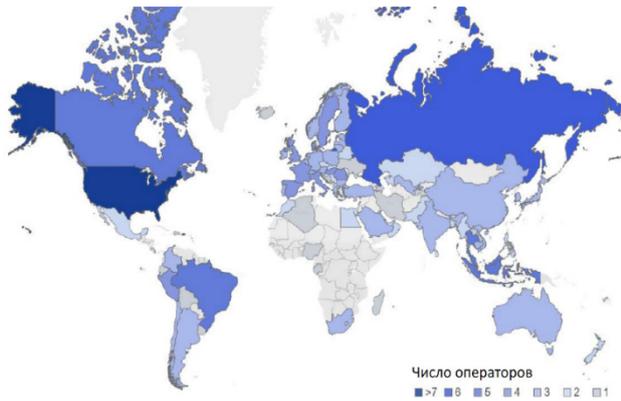**Figure 6.** Number of 5G-NR base stations in commercial circulation [3]

Figure 7. Commercial, planned and pilot 5G networks [6]

Figure 8. Scenarios for building a 5G network in the initial and intermediate periods [6]

As for subscriber equipment, there are already many models of 5G smartphones, routers and CPEs on sale.

Early adopters are already seeing a significant increase in 5G transmission speeds. Qualcomm test results (May 2019) show a 3.3x increase in download speed for 5G devices compared to LTE devices. In the future, this figure will be higher due to denser coverage and the transition from the LTE EPC core to the packet 5G network core.

THE PATH FROM LTE TO 5G

As mentioned, current 4G networks are not up to the demands of new application scenarios. In the long term, full-fledged 5G networks, including those to support Network Slicing and URLLC, will require a new NGCN (Next Generation Converged Network) network infrastructure, as well as an upgrade of the radio access network. It is clear that it is impossible to perform such a volume of work at once.

4G network delays

The 3GPP consortium initially took into account the complexity of deploying new networks and adopted scenarios for moving from the standard configuration of LTE networks (# 1) to 5G. The introduction of 5G is proposed to be carried out first over the existing LTE EPC infrastructure in NSA (Non-Standalone, #3) mode, as cellular operators did throughout 2019. In such a configuration, the delays on the radio part will be reduced, but due to the restrictions of the EPC core LTE, the overall delay will be far from the requirements of the URLLC. The main meaning of this configuration is different – in the radio part we will get a significant increase in throughput sufficient for most existing eMBB applications, as well as connection stability with a large number of connected subscribers per base station (Fig. 8).

NSA's initial model (#3) aims to improve the quality of mobile broadband internet to improve reliability and data capacity by using EN-DC (E-UTRAN New Radio – Dual Connectivity) connectivity.
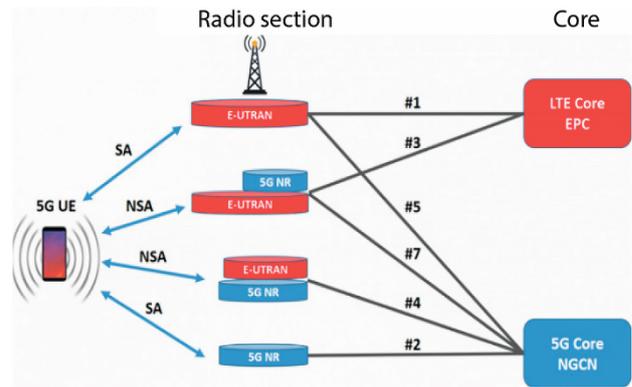
User terminals that support EN-DC can simultaneously connect to LTE and 5G base stations, while the LTE base station is an anchor (upgrading to ng-eNb, or new generation eNB is required). The user terminal (UE) initially registers with the network via E-UTRAN at low frequencies (<2 GHz) and starts broadcasting the measurements performed on the 5G-NR radio access network to the network. When the 5G "radio quality" is satisfactory, the LTE ng-eNb base station initiates a request to the 5G gNB base station to allocate network resources to the UE.

Upon completion of the procedure, the UE connects simultaneously to the LTE ng-eNB and 5G gNB base stations. Of course, the coverage area of a 5G base station will be much narrower than LTE. a high-frequency millimeter-wave signal has a higher attenuation coefficient (Fig. 9).

Figure 9. UE connectivity to LTE ng-eNB and 5G gNB in EN-DC mode [6]

Further, through combined LTE + 5G-NR base stations, 5G coverage can be expanded by using DSS (Dynamic Spectrum Sharing, dynamic spectrum sharing) technology, when the lower frequency band of E-UTRAN (<2 GHz) is dynamically shared with 5G-NR . Prior to the introduction by operators of the 5G core, networks will be able to work like this (Fig. 10).
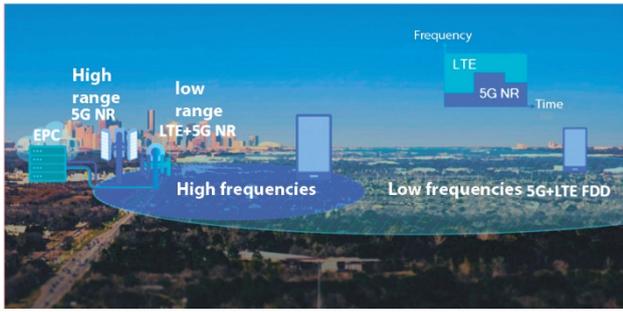
Figure 10. Expanding 5G Coverage with Low Frequency LTE (DSS) [6]

Figure 12. Expansion of the coverage area of medium frequencies due to an unpaired channel [6]

Further from step #3, when cellular operators integrate the 5G NGCN core, they can move to the target and final SA mode (options #2 and #5) when using one radio access technology – either E-UTRAN or 5G-NR. Below is the final view of the 5G network capable of providing URLLC services. (Fig. 11).
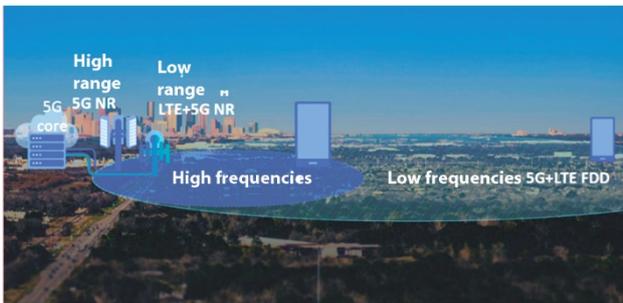

Figure 11. Final view of 5G network [6]

To meet the growing demand for eMBB, medium frequencies (2 GHz-7 GHz) can be used, thus increasing data rates, including through frequency aggregation. Lower frequency - more coverage, but less channel width. However, there is a way to increase coverage while maintaining a high upload speed using an additional uplink channel (SUL, Supplementary Uplink). How it works? The figure below shows how a UE's Medium Frequency Paired (UL/DL) radio resource is assigned a Supplemental Low Frequency Unpaired Uplink Channel (SUL). Then, in one cell, the UE receives 1xDL (mid frequencies) and 2xUL (low and medium frequencies) channels, the use of which will be controlled by the network. In this case, at the cell boundary, the DL channel uses a medium-frequency signal with increased power from the "paired" band, and the UL channel uses a low-frequency signal in the unpaired SUL band. As a result, the base station "sees" the UE at longer distances, and the download speed is maintained as with the use of medium frequencies (Fig. 12).

5G COVERAGE FORECAST

From the 5G network deployment scenarios and applied frequency bands, the logic of frequency band allocation for different scenarios follows, as shown in the table. Subject to this concept is the Network Slicing technology defined by the 3GPP specification; it will allow mobile operators to deploy networks isolated from each other, each of them can be allocated for specific needs (for the Internet of things, streaming video, etc.) (Tabl. 3).

Table 3

Application scenarios in different 5G frequency bands

Frequencies	Bandwidth	Scenarios	C acteristic
above 7 GHz (FR2)	800 MHz	eMBB	Super high speed, small coverage and only on the streets
2 GHz … 7 ГГц (FR1)	100 MHz	eMBB, URLLC, mMTC	High speed, wide street coverage, satisfactory indoor coverage
< 2 GHz (FR1)	20 MHz	eMBB, URLLC, mMTC	Medium speed, ubiquitous street and indoor coverage

Given the feasibility of these scenarios in megacities, small towns and villages, it is possible to draw up a generalized network coverage scheme as in Figure 34. As you know, the world has either already abandoned 3G networks or has already announced plans to turn them off. Therefore, Figure 3G is shown as a dashed line. According to the plans of European operators, instead of slow 3G, fast LTE will come to the villages (scenario #5) while maintaining 2G for voice traffic. Suburban coverage will feature faster speeds and lower latency, while metropolitan areas will also boast more connections and ultra-fast internet in FR2 coverage areas. As you can see, in the coming years, 5G networks will not replace 4G networks, but will gradually integrate with 4G networks, significantly improving the overall situation (Fig. 13).
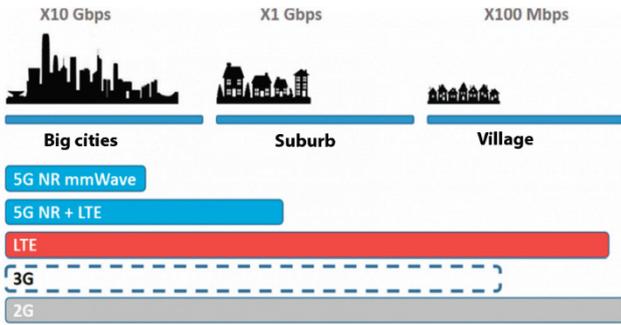
Figure 13. Generalized scheme of coverage
by 2G, LTE and 5G networks until 2025

Separately, it must be said that such a distribution of networks will give rise to a sharp growth in the FWA (Fixed Wireless Access) market. Manufacturers of CPE (Customer Premises Equipment, wall-mounted or in-house telecommunications equipment located in the subscriber's premises) will be able to provide high-speed Internet to residents of territories where high-speed and reliable 5G coverage for some reason "did not reach".

Typically, wired and fiber optic Internet providers come to such places. But 5G FWA will pose a serious threat to this business sector. Indeed, with 5G, FWA's broadband Internet quality will not be inferior to fiber-optic Internet, and the cost of connection will be completely out of competition, since fiber/cable installation, installation work and configuration for each subscriber individually are excluded. Installing the CPE is as easy as installing a Wi-Fi router and requires almost no maintenance (Fig. 14).
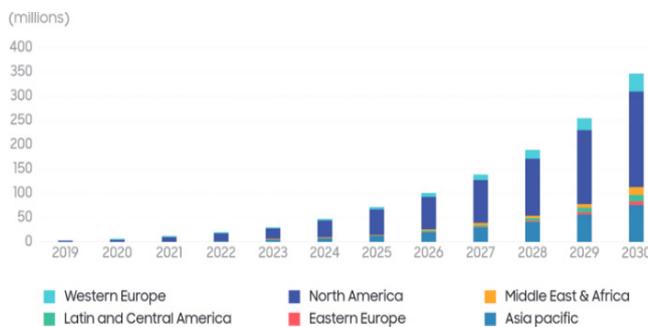

Figure 14. Number of FWA connections using 5G

Source: SNS Telecom

Perhaps, as a result, the fiber optic/wired Internet market will be greatly transformed, will find its specific application, but will never be as massive as it is today. On the contrary, according to the forecast of experts from SNS Telecom (www.snstelecom.com), by 2030, 345 million subscribers will be connected via the FWA service via 5G, and over 90 million units of CPE subscriber devices will be sold. In Russia, this service, due to the length of the territory, can be in great demand even at the initial stage of deploying 5G networks.

5G & GNSS

The virtual Primary Reference Time Clock (vPRTC) timing architecture uses a secure fiber network to provide bi-directional east and west timing flows to local source clocks in aggregation sites with 100 ns of accuracy. The protected PTP southbound timing flow is then sent to the 5G Remote Radio Heads (RRHs). A key benefit of this architecture is that the GNSS receivers are only required in the core enhanced Primary Reference Timing Clock (ePRTC) sites where they can be protected and monitored to shield the network from the risk of GNSS jamming or spoofing threats [7].

To support secure 5G network communications, the vPRTC is a highly secure and resilient network-based timing architecture that blends our secure GNSS firewall technology, high-precision atomic system clocks and a portfolio of trusted time solutions to meet the expanding needs for 5G deployment [21].

The combination of GNSS visibility and the vPRTC architecture provides 5G operators with a dual-purpose solution for the use of Positioning, Navigation and Timing (PNT) services as delivered by GNSS. It provides situational awareness about the health of GNSS reception by monitoring and evaluating key GNSS observables in real time to determine if there is risk in the use of PNT delivered by GNSS. Like a network firewall, this solution creates a dome of protection that strengthens the overall use of PNT services by GNSS for critical 5G communication infrastructures [22].

Prior generation wireless networks relied heavily on the Global Navigation Satellite System (GNSS) to provide a traceable time reference for all the source clocks in the Radio Access Network (RAN). Well-engineered GNSS-based clocks can easily provide +/- 100 ns Primary Reference Time Clock (PRTC) timing accuracy. Table 4 shows the timing accuracy for common source clocks as defined by the ITU for use in communications networks.

Table 4

Primary Reference Time Clock (PRTC) types
available for 5G [7]

Clock Type	Accuracy	ITU standard	Comment
PRTC – A	100 ns	ITU-T G.8272	Single band GNSS
PRTC – B	40 ns	ITU-T G.8272	Typically, dual-band GNSS
Enhance PRTC (ePRTC)	30 ns	ITU-T G.8272.1	Combines GNSS with a local Cesium Atomic Clock at core sites to provide an autonomous timescale that can maintain timing if GNSS is lost

5G Challenge with GNSS

5G operators face two major challenges with deployment of GNSS-based source clocks.

First, 5G networks are very dense, often requiring ten times as many source clocks compared to prior generation networks. That means ten times as many GNSS antennas to install and maintain. GNSS antennas are already installed in core/RAN sites, but not out in 5G aggregations sites. GNSS antennas are a headache you would rather avoid.

Second, GNSS is vulnerable to cybersecurity attacks such as jamming and spoofing. Your 5G radios have a minimum holdover time and will quickly begin interfering with other radios in the same spectrum if GNSS is compromised due to failures or other vulnerabilities such as jamming.

Are GNSS Cybersecurity Threats Real? In a word – yes. And, government agencies are finally stepping forward to do something about it [7]. If you don't believe GNSS threats are real, spend a few minutes searching them out on the internet. The below table shows some of the recent government actions to address this cybersecurity threat (Tabl. 5).

Table 5

Government initiatives to address GNSS cybersecurity

Action	
Executive order on strengthening national resilience through responsible use of Position Navigation and Timing (PNT) services	February 2020
US National Space Policy – Key initiatives to improve Cybersecurity for GNSS systems	December 2020
Department of Homeland Security (DHS) Resilient PNT Conformance Framework	December 2020

For example, Microchip's innovative new timing architecture – the virtual Primary Reference Time Clock (vPRTC) – uses protected core ePRTC timing sites as redundant area timing hubs for your network. The ePRTC achieves 30 ns accuracy when locked to GNSS and can maintain 100 ns accuracy for a minimum of 14 days if GNSS is lost. This architecture helps to deliver a new resiliency – which in turn helps network operators provide uninterrupted services to their clients (Fig. 15).
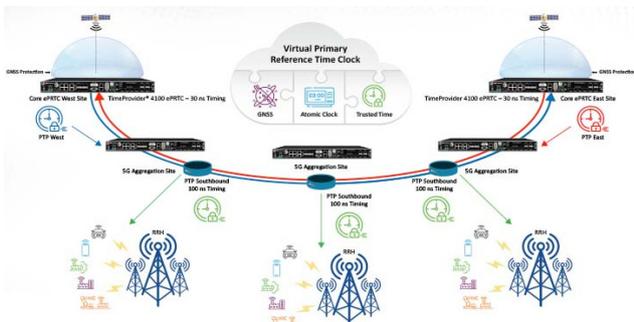

Figure 15. Bi-directional protected timing flows deliver required 5G timing accuracy without the need for GNSS at 5G aggregation sites [7]

The above diagram shows how redundant core ePRTC sites can be deployed for a 5G network. Bidirectional protected Precision Time Protocol (PTP) timing flows from the "west" and "east" core ePRTC sites to all 5G aggregation sites. Operators and optical equipment suppliers have presented technical papers on similar timing architectures at the International Timing and Synchronization Forum (ITSF) conference in November 2021 demonstrating field results supporting timing accuracy and resiliency [2-4].

Microchip's TimeProvider 4100 clocks at 5G aggregation sites are configured in High Performance Boundary Clock (HPBC) mode to receive timing from the core "east" and "west" sites with an error budget of less than 5 ns. If timing flow from the west is lost, the clock will instantly switch to the protected east flow to maintain required timing accuracy for the network without the need for a GNSS receiver at the aggregation sites.

The southbound PTP flow out to the RRHs is protected by the bi-directional timing flows from the core east and west ePRTC sites. Your core timing sites can also be protected with the BlueSky™ GNSS Firewall technology to monitor GNSS observables, detect potential jamming or spoofing threats and shield them from impacting core ePRTC sites.

CONCLUSION

The cost of passing ever-increasing traffic through the networks of telecom operators as of 2020 is not covered by income from traditional services. The search for new services, the so-called "killer application" of traditional telecom platforms, usually does not give the expected results. Meanwhile, the main growth in traffic and revenues is not in the human device sector, but in the IoT device sector, which is one of the basic goals of 5G functionality. Therefore, 5G networks can be considered one of the necessary components of digital transformation and the digital economy.

Unmanned vehicles can act as part of the Smart City service, however, they can be provided on their own platform. It includes not only driverless cars, but also unmanned tractors for Smart Agriculture, unmanned subway and commuter trains, drones and other types of public and special transport. In addition, the implementation of ADAS (Advanced Driver-Assistance Systems) driver assistance systems is possible on the 5G platform.

Unlike networks of previous generations, the range of services of which was severely limited and somewhat expanded in 4G, the services of the 5G platform are synergistic and scalable, and are not limited to once set functionality. 5G plays the role of a platform for the development mode of new DevOps services and applications, when new features are created by developers in close coordination with the teams responsible for their implementation and operation.

5G will not necessarily change the risk factors we have. But it will exponentially increase the threat vectors and opportunities that attackers can take advantage of.

In general, 5G network includes not only mobile, but also fixed communication services, as well as high-speed Internet access with low latency and specialized corporate networks for vertical sectors of the economy.

Due to fifth generation networks, it will also be possible to improve the quality of use of existing services, where large amounts of traffic are involved. The 5G network platform provides operators with significant benefits, primarily expressed in expanding the functionality and characteristics of the network and improving the User Experience [23, 24].

It is still too early to say that 5G networks have acquired their mature, final form. We need to wait for the 16th release, which, according to the consortium's plan, will close the second phase of the specifications and determine the beginning of the mass introduction of fifth-generation core networks. However, this does not prevent from starting work on the study of new technology today, which will lay the foundation for future projects. 5G networks will become our daily routine and the transition from NSA to SA mode will be smooth and imperceptible, and the achievements made today will not be wasted.

REFERENCES

1. Department of Homeland Security; Resilient Positioning, Navigation, and Timing (PNT) Conformance Framework Version 1.0. December 2020.

2. Helmut Imlau (2021). Resilience for Timing & Synchronization Networks, *International Timing and Synchronization Forum*, November 2021.

3. Helmut Fabian (2021). Implementation of Reliable Synchronization for 5G, *International Timing & Synchronization Forum*, November 2021.

4. Jon Baldry, Infinera (2021). Key Enablers for Migration to Timing Cloud Synchronization Distribution in Optical Transport Networks, *International Timing & Synchronization Forum*, November 2021.)

5. Batuev B.B. (2017). SIM7000E/SIM7000E-N: application of power-saving PSM and eDRX modes in the NB-IoT network. *Wireless Technologies*. No. 2.

6. Why and how 5G will change everything: technologies, phased implementation and element base for subscriber equipment. https://habr.com.

7. Barry Dropping (2020). Do I Really Need GNSS Timing Receivers in Every 5G Aggregation Site? Microchip.com.

8. What is 5G? How will it transform our world? https://ericsson.com/5G.

9. 5G technologies will transform your network. https://nokia.com/networks/5G.

10. 5G Solutions. https://www.keysight.com/us/en/solutions/5g.html.November 8, 2018.

11. https://academy.itu.int.

12. Tikhvinsky V.O. (2017). The fifth element of the mobile world: the results of MWC-17. *T-Comm*. 2017. Vol. 11. No. 3, pp. 4-11.

13. Fokin G.A. (2020). 5G network positioning and statistic models for its accuracy evaluation. *T-Comm*, vol. 14, no.12, pp. 4-17.

14. Glazkov R.V., Nikitina A.V. (2019). Heterogeneous networks spectral efficiency analysis with modified time-domain interference coordination algorithm in various load distribution scenarios for 5G New Radio. *T-Comm*, vol. 13, no.12, pp. 56-61.

15. Pankratov D.Yu., Stepanova A.G. (2018). Computer simulation of MIMO technology for radio systems. *T-Comm*, vol. 12, no.12, pp. 33-37.

16. Vorozhishchev I.V., Bochechka G.S. (2017). Sensitivity analysis of UFMC to carrier frequency offset estimation errors. *T-Comm*, vol. 11, no.6, pp. 25-28.

17. Krukov Ya.V., Demidov A.Ya., Pokamestov D.A. (2018). Multiple access method with power division channels on orthogonal frequencies. *T-Comm*, vol. 12, no.1, pp. 17-22.

18. Timoshenko A.G., Molenkamp K.M., Molenkamp N.B., Belousov E.O. (2017). The influence of technological mismatch on synchronization system behaviour with Zadoff-Chu sequences. *T-Comm*, vol. 11, no.12, pp. 8-13.

19. Umer Mukhtar Andrabi, Stepanov S.N., Ndayikunda J., Kanishcheva M.G. (2020). Cellular network resource distribution methods for the joint servicing of real-time multiservice traffic and grouped IoT traffic. *T-Comm*, vol. 14, no.10, pp. 61-69.

20. Stepanets I.V., Fokin G.A., Mueller A. (2018). Capacity estimation ways of massive MIMO systems. *T-Comm*, vol. 12, no.10, pp. 64-69.

21. Koval V.A. (2018). Development of institutional/economical facility of radio-frequency spectrum conversion management for improving the revenue of mobile communication services. *T-Comm*, vol. 12, no.6, pp. 61-64

22. Elagin V.S., Belozertsev I.A., Onufrienko A.V. (2019). Models of QOE ensuring for OTT services. *T-Comm*, vol. 13, no.4, pp. 9-14.

23. Konstantinov N.E., Gorodnichev M.G., Gematudinov R.A. (2018). Blockchain as an IoT development platform. *T-Comm*, vol. 12, no.9, pp. 63-68.

24. Bakulin M.G., Kreyndelin V.B. (2020). The problem of spectral efficiency and capacity increase in perspective 6G communication systems. *T-Comm*, vol. 14, no.2, pp. 25-31.

25. Kreindelin V.B., Usachev V.A. (2017). LTE-Advanced pro as a basis for new use case M2M. *T-Comm*, vol. 11, no.3, pp. 28-32.

26. Kreiyndelin V.B., Smirnov A.E., Ben Rejeb T.B.K. (2016). Efficiency of signal processing in multiuser large scale MIMO systems. *T-Comm*. Vol. 10. No.12, pp. 24-30.

27. Kreindelin V.B., Starovoytov M.Yu. (2017). Application of predetector receive signal combining to improve the performance of MIMO spatial multiplexing. *T-Comm*, vol. 11, no.4, pp. 4-13.

28. Kreyndelin V.B., Grigorieva E.D. (2019). The implementation of the bank of digital filters with reduced computational complexity. *T-Comm*, vol. 13, no.7, pp. 48-53.

29. Kreyndelin V.B., Reznev A.A. (2018). Golden type high dimensional matrix for space time coding. *T-Comm*, vol. 12, no.6, pp. 34-40.

30. Kreyndelin V.B., Grigorieva E.D. (2019). Modification of bilinear transformation method and its application to digital filters synthesis. *T-Comm*, vol. 13, no.1, pp. 4-9.