Informacije MIDEM

Informacije MIDEM *2-2013*
Journal of Microelectronics, Electronic Components and Materials

Content | Vsebina

Editorial | Uvodnik

Dear Reader,

For all those who have not paid attention yet to the Table of content or the upper left corner of each paper published in renewed outlook I would like to point out that all papers are classified in three categories:
- Review scientific paper,
- Original scientific paper, and
- Professional article.

Last year (volume 42 with its four issues) we have publish 34 original scientific papers and one professional paper. In the current issue it is my pleasure to announce a publication of the first review scientific paper that covers the research field of RF MEMS electrostatic switches.

I hope the review paper will encourage also you to submit a review paper in your field of expertise. In preparation of your manuscript do not forget to search for relevant papers published in our Journal. You can use a search engine on the web page http://www.midem-drustvo.si/Journal/Home.aspx, where you just need to click "Advanced paper search" by title or author.

We are proud to announce another improvement related to the visibility of our Journal. Recently we have become a member of vast **Google Scholar** database with all published papers in Informacije MIDEM since 1989!

I am sure that this improvement will increase the readers of our journal followed by wider recognition in the field of microelectronics, electronic components and materials.

Marko Topič
Editor-in-Chief

P.S.
All papers published in Informacije MIDEM (since 1986) can be access electronically at
http://midem-drustvo.si/journal/home.aspx
or
Google Scholar.

Review of low actuation voltage RF MEMS electrostatic switches based on metallic and carbon alloys

Yasser Mafinejad[1], Abbas Kouzani[2], Khalil Mafinezhad[3]

[1]School of Engineering, Deakin University, Waurn Ponds, Australia
[2]School of Engineering, Deakin University, Waurn Ponds, Australia
[3]School of Electrical Engineering, Sadjad Institute of Higher Education, Ferdowsi University, Mashad, Iran

Abstract: Radio frequency micro electro mechanical systems (RF MEMS) have enabled a new generation of devices that bring many advantages due to their very high performances. There are many incentives for the integration of the RF MEMS switches and electronic devices on the same chip. However, the high actuation voltage of RF MEMS switches compared to electronic devices poses a major problem. By reducing the actuation voltage of the RF MEMS switch, it is possible to integrate it into current electronic devices. Lowering the actuation voltage will have an impact on RF parameters of the RF MEMS switches. This investigation focuses on recent progress in reducing the actuation voltage with an emphasis on a modular approach that gives acceptable design parameters. A number of rules that should be considered in design and fabrication of low actuation RF MEMS switches are suggested.

Keywords: RF MEMS switch, low actuation voltage, RF parameters, metallic and carbon alloys

Pregled RF MEMS elektrostatičnih stikal z nizko napetostjo vzbujanja na osnovi kovinskih in karbonskih zlitin

Izvleček: Mikro elektromehanski sistemi na radijski frekvenci (RF MEMS) so omogočili razvoj novih naprav predvsem zaradi njihove visoke učinkovitosti. Obstajajo številne spodbude k uporabo RF MEMS stikal in elektronskih naprav na enem samem čipu, vendar visoka vzbujevalna napetost predstavlja problem. Z znižanjem vzbujevalne napetosti lahko RF MEMS stikala vgradimo v obstoječe naprave. Znižanje napetosti pa bo imelo velik vpliv na RF lastnosti stikal. V članku so predstavljene trenutne smernice nižanja vzbujevalne napetosti na osnovi modularnega pristopa, ki omogoča sprejemljive parametre načrtovanja. Predlagana so številna pravila, ki bi se naj upoštevala pri načrtovanju in izdelavi RF MEMS stikal.

Ključne besede: RF MEMS stikala, nizka napetost vzbujanja, RF parametri, kovinske in karbonske zlitine

* Corresponding Author's e-mail: ymafinej@deakin.edu.au

1 Introduction

Micro electro mechanical systems (MEMS) have enabled a new generation of electronic devices, particularly RF switches. MEMS switches can be employed in radio frequency (RF) circuits, and their performances could be made better than those of other standard switches such as FET, and PIN diodes [1]. This is due to their good linearity, low noise, low power consumption, high electrical isolation, and ultra wide frequency band [2]. MEMS switches are designed to operate for DC to a few hundred GHz applications [3, 4]. For example, they can be used in cell phones, short range communication systems such WLAN and Bluetooth [5, 6], automotive systems such as acceleration and gyro sensors [7, 8], biomedical devices such as lab-on-a-chip [9-12], and radar applications [13, 14].

RF MEMS switches can be categorized into four groups according to their types of actuation forces. The first

type is the Piezoelectric RF MEMS switches. This type of switch uses piezoelectric materials such as AlN or PZT on top of the membrane or beam. These piezoelectric materials cause an elongation and strain across the length of the piezoelectric layer and make the beam deflect by applying the voltage. The amount of force depends on the piezoelectric coefficient. Therefore, the low actuation voltage can be achieved by chosen a high piezoelectric coefficient [15-23]. The second type of MEMS switch is the electromagnetic RF MEMS switch. This type of switch uses coil on top of the membrane. The electromagnetic force is created when a DC current is applied to the coil and actuates the membrane [24-26]. The third type is the electro-thermal RF MEMS switch. The bending of the structure depends on the thermal expansion coefficient of materials.Applying a current through the resistor on top of the beam causes a thermal wave propagates and attenuates in the thickness direction and it deflects the beam. [27, 28]. The last and more applicable type is the electrostatic RF MEMS switch. This type of switch operates only based on the amount of actuation voltage and the capacitance between the transmission line and the membrane. Table 1 compares all types of the RF MEMS switch. As can be seen from the Table 1, electrostatic force performs better in all parameters except for the actuation voltage, which is very high. Although the switching time and reliability of the electrostatic MEMS switches are better than those of other types, they not compare well with other RF switches, such as semiconductor and mechanical switches [29-33].

Carbon allotropes such as graphene and carbon nano tubes have shown superior electrical and mechanical performance compared to other types of materials. Nowadays, carbon allotropes are used widely in microwaves [34, 35]. Due to the great advantage of these materials, they can be a very good candidates for development of the RF MEMS switches [36].

In the future, electronic industry will need to integrate the MEMS and electronic devices on the same substrate. Our experience with the nonlinearity and noise of amplifiers shows that applying MEMS technologies for RF components such as RF tuning filter, switches, phase shifter, and transceiver systems will provide a far better performance than the current techniques [37-42]. Moreover, this technology also provide advantages such as miniaturizing the size, enhancing signal transduction, reduced chip pin out, increasing immunity to the electromagnetic interference, reducing power loss, and offering lower cost compared with multichip implementations [43-45]. The main problem for integration of the RF MEMS electrostatic switches with the current electronic devices is their very high actuation voltage (more than 50 Volts).

Although there are some review papers and books covering RF MEMS switches, there are as yet no book or paper discussing methods developed for lowering the actuation voltage of the RF MEMS switches. The aim of this investigation is to introduce, explain and compare recent techniques and materials developed for lowering the actuation voltage of the electrostatic RF MEMS switches that can be used in IC technology (<15V).

Section 2 discusses the principles of the RF MEMS switches based on metals and carbon materials, and their main parameters. Section 3 focuses on recent methods used for lowering the actuation voltage and compares them and their effects on switch parameters.

Table 1: Comparison of different types of MEMS switch actuation

	Piezo-electric	Electro-thermal	Electro-magnetic	Electro-static
Size	Medium	Medium	Large	Small
Fabrication process	Complex	Medium	Complex	Simple
Actuation voltage	Medium	Low	Low	High
Power con-sumption	Medium	High	High	Low
Switching speed	Fast	Slow	Medium	Fast
Reliability	Medium	Low	Medium	High

2 Principles of MEMS switches

In order to provide a background to the RF MEMS switches, this section presents basic information on the structure, fabrication, modelling, and categorization of such switches.

2.1 MEMS switch structure

MEMS switches have different shapes. All MEMS switches have the same structure, which consists of three components [46-49], as shown in Figure 1. The first component is substrate, which is the basic element of any microelectronic device, and is used to mount a device on top of it (surface micromachining), or inside it (bulk micro machining). In order to integrate the RF MEMS switch with IC circuits, high resistivity silicon such as HR or porous silicon [50-52], is used. This is due to the high loss of normal silicon at high frequencies. The substrates that can be used in RF MEMS switches are given in Table 2. The second component is a transmission line, which is used for transferring the RF signals from the input to the output ports. The third

component is a cantilever or membrane, which are movable parts for connecting and disconnecting the signal line (dual-fixed bridge shunt switch) (Figure 1).
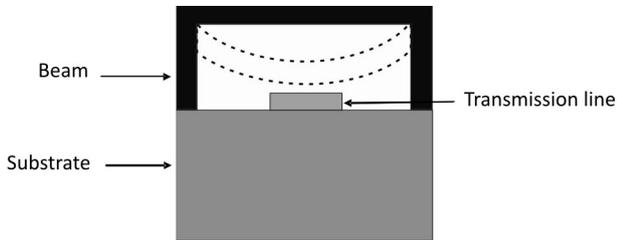
Figure 1: Structure of RF MEMS switch (dual-fixed bridge shunt switch)

Table 2: Materials used for substrate

Substrate	Quartz	Alumina (Al)	Sapphire	Silicon	Galium arsenide
Relative Permittivity	3.78	9.75	11.72	11.72	12.91

2.2 Materials of RF MEMS switches

RF MEMS electrostatic switches can be categorized, based on their materials, into two groups. The first group is the metallic MEMS switch, which uses metallic alloys such as copper, aluminium and gold. The second group is the carbon MEMS switch, based on carbon nano tube (CNT) or graphene. CNT is a type of carbon allotrope where rolled sheets of sp2-bonded graphene are shaped into a long hollow tube. CNT can be categorized into a single- wall carbon nanotube (SWCNT) and multi-wall carbon nanotube (MWCNT). Graphene is a type of carbon allotrope which can be geometrically considered as single atomic layer of carbon. The electrical and mechanical properties of these materials are superior to other materials (Table 3). Also, these materials reduce the size of switch from micro to nano meter (NEMS switches) [53-57].

Table 3: Materials for MEMS and NEMS switches

Types of switch	Material	Resistance μΩ x cm (ρ)	Young's modulus (GPa)
MEMS metallic switch	Copper	1.69	117
	Gold	2.2	79
	Aluminium	2.65	69
MEMS carbon switch	CNT	10	1000
	Graphene	10	1000

2.3 Fabrication of MEMS switches

MEMS switches based on metal are fabricated using two techniques. The first technique is bulk micromachining, which is based on the etching of silicon substrate and it relies on the etch rate of the crystal direction [58]. The second technique is surface micromachining. It is based on lithography, deposition of metals and etching of sacrificial layers to release the bridge or membrane on the transmission line, and uses 5-6 masks. More information on metal-based switches fabrication is given in [59-63].

Fabrication of the RF MEMS switches based on carbon material can be achieved using three steps as described in [64, 65]. The first step is to pattern the transmission line. The second step is to grow CNT or Graphene by chemical vapour deposition [66]. The third step is to pattern metal contacts onto two edges of the beam.

2.4 MEMS model

Analysing the RF MEMS switches requires extracting the mechanical and electrical models of the switches.

Mechanical model
A comprehensive study of the dynamics and statics of MEMS switches can be found in [67-69] . Figure 2 shows the mechanical model of the RF MEMS switch. There are three types of forces involved in MEMS switches. First, the Van der Waals force, which plays an important role while the gap between the two electrodes is in the range of a few nano meters. The second force is the electrostatic force, which relies on a voltage source and a capacitor between the TL and the membrane. The third force is due to the elastic force, which is modelled as a spring and depends on the shape, material and size of the beam or the membrane.
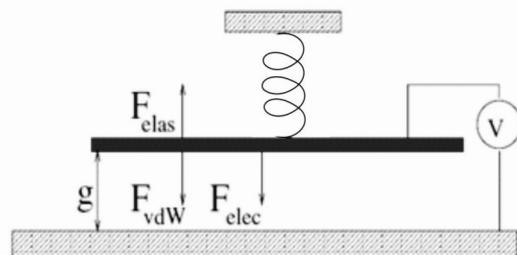
Figure 2: Mechanical model of a MEMS switch

Actuation voltage makes instability and causes the upper electrode to snap down. Another important mechanical parameter of the MEMS switch is the switching time. The switching time of MEMS switches is limited by the mechanical structure. The pull in actuation voltage ($V_{pull-in}$) and switching time (t_s) for vertical types of MEMS is as follows [70-73]:

$$V_{pull-in} = \sqrt{\frac{8kg_0^3}{27A\varepsilon_0}}$$

(1)

$$t_s = 0.46 f^{-1} \tag{2}$$

$$f = \sqrt{\frac{k}{m}} \tag{3}$$

where k is the spring constant, g_0 is the gap between electrodes without actuation voltage, A is the overlap area between the bridge and the transmission line or the electrode, f, m is the mass of beam and f is the first resonant frequency of the beam.

Electrical model

The switch has two states: On and Off. The RF parameters of the switch, such as S_{11} and S_{21}, can be calculated by the electrical model in both states [74] . Figure 3 shows the RF MEMS shunt and series switches, which are modelled by electrical circuits. The switch is modelled by C, L, and R components. L represents the inductance of the switch, R shows the insertion loss, and C, which is the dominating parameter, represents the capacitance between the bridge and the transmission line. This capacitance has two extreme values at the up state and the down state and varies between them. The values of S_{11} and S_{21} strongly depend on the capacitance of the bridge. For example, the amount of S_{11} and S_{21} for the shunt switches are given by (4-6)

$$\left|s_{11}\right|^2 + \left|s_{21}\right|^2 = 1 \tag{4}$$

$$s_{11}(up\ state) = \frac{-j\omega\, C_{up} Z_0}{2 + j\omega\, Z_0} \tag{5}$$

$$S_{21}(down\ state) = \frac{2}{2 + j\omega\, C_{down} Z_0} \tag{6}$$

According to equations 1-5, table 4 summarizes the impact of physical parameters on both mechanical and electrical properties of the switch. For example, reducing the gap between the signal line and the bridge (g) reduces the actuation voltage. However, this increases the up state capacitance (C_{up}) and diminishes the isolation (S_{21}). Therefore, it reduces the bandwidth at up state (S_{21}). Moreover, although reduction of spring constant (K) reduces the actuation voltage but it reduces the resonant frequency or increasing the switching time.
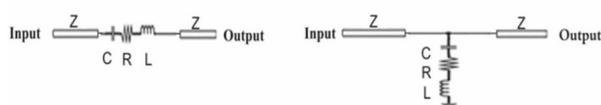
Figure 3: Series and shunt electrical model of MEMS switches

Table 4: Effects of physical parameters on the mechanical and RF properties of the switch

Physical parameters	Mechanical properties		Electrical properties	
	Vpull-in	Frequency Resonant	Up state bandwidth	Down state
Reduction of g	Reduce	Does not effect	Reduce	Does not effect
Reduction of K	Reduce	Reduce	Does not effect	Does not effect
Increase of size	Reduce	Reduce	Reduce	Increase

3 Review of low actuation voltage RF MEMS switches

As was mentioned in the previous section, MEMS switches have different parameters which should be considered in their design and fabrication. This section reviews the methods that have been used to reduce the actuation voltage while considering the requirements of other parameters of the switches as described in previous section. The methods used to reduce the actuation voltage of switches can be categorized into three groups.

3.1 Low actuation voltage based on reducing the gap

According to Table 4, reduction of the gap decreases the actuation voltage but it also deteriorates the RF parameters. The following techniques are used to reduce the actuation voltage while maintaining RF parameters at acceptable levels:

3.1.1 Matching circuit

The amount of parasitic capacitance affects negatively on the insertion loss of the MEMS switches at up state position. One effective method for compensating this capacitance is by using a matching circuit accompanied by a switch. In this way, the amount of gap can be reduced at any range. Mafinejad et al. [75-79] reported a low actuation voltage shunt capacitive contact for the frequency band of Ka to V, by using a T and π matching circuits. The T match circuit uses two short high impedance transmission lines (SHITL) before and after the switch on CPW transmission line while a π match was used only one SHITL between two switches on CPW. The SHITL can be achieved by increasing the distance between TL and the ground or narrowing the signal line of CPW (Figure 4).
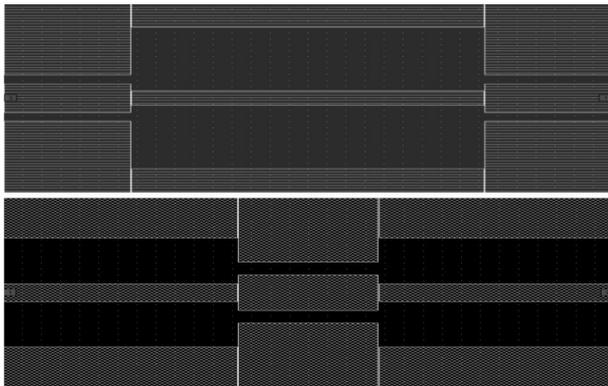
Figure 4: π and T matching circuits

3.1.2 Using pillars and an extra voltage source

The second method for reducing the actuation voltage is by using the structure which is shown in Figure 5. It consists of a beam which is anchored to a pillar at the middle whilst leaving the ends free. This method uses two separate voltage sources to provide a negative and positive voltage. This type of switch has three states: On, Off and Neutral. When voltages are applied to either of the electrodes, a large positive deflection is then noticed on one side and a smaller negative deflection of the membrane is obtained with a large contact area on the other side. This creates a large capacitance ratio for up and down state position.
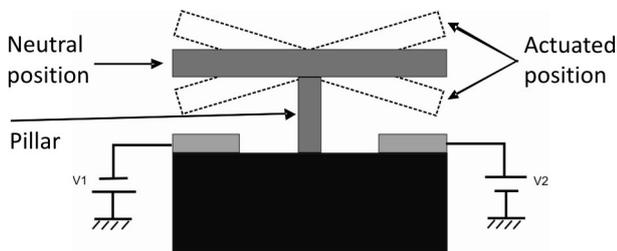
Figure 5: RF MEMS switch by using pillars and two voltage sources

Tauati et al. [80] designed a low actuation voltage (5 v) series switch that can be used for DC up to 10GHz. The switch used two pillars and four electrodes (two internal and two external). Robin et al. [81] proposed a RF MEMS SPDT switch with an actuation voltage of 20V for the frequency band of 15-30GHz. The switch consists of three pillars for the support of a gold membrane, and four electrodes (two internal and two external). Kim et al. [82] designed a Single Input Double Output (SIDO) switch from 2-10 GHz application . The actuation voltage of this switch is 15V. The switch consists of dual fixed beam and a pillar which is positioned under the beam leaving a small gap. When the SW2 changes to the ON state, SW1 is restored by the leveraging force as well as by its own stiffness. After SW2 is turned ON, the SW1 is forced to maintain higher bending stiffness

against the self-actuation power with the help of an axial force and leveraging moment.

3.1.3 Comb switches

Unlike the MEMS switches which have vertical actuation, this type of switch has a lateral actuation. The lateral switch consists of three main parts: A comb driver, which consists of stationary and movable combs to provide electrostatic force (Figure 6). The second part is a flexible structure which works as a beam or membrane of vertical switches and is connected to the driver. The third part consists of a transmission line [83-85].
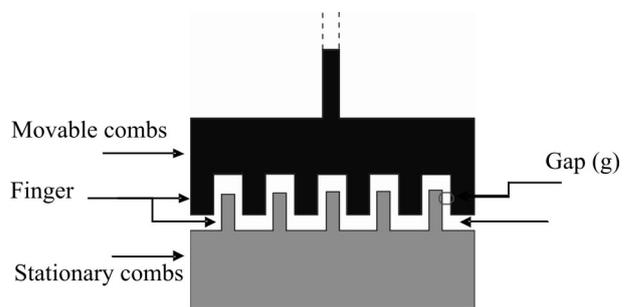
Figure 6: Comb driver

The electrostatic force (F_{es}) between two moveable and stationary combs as a result of applied voltage between them is given as:

$$F_{es} = \frac{\varepsilon_0 t}{g} NV^2 + \frac{\varepsilon_0 wh}{d} NV^2 \qquad (7)$$

g is the gap, t is the thickness of the comb drive, d is the gap between the moveable and stationary combs, N is the number of fingers, and d is the distance between movable and stationary fingers.

Kang et al. [86] designed and fabricated Single Input 4 Output (SI4O) and SI12O RF MEMS series DC contact switches. The actuation voltage of both switches is 15V and the frequency band is DC to 10 GHz. The gap between the flexible structure and the transmission line is 2.5μm and the actuation voltage is 15V. The switching speed is 120μs and 500μs for the switch ON and OFF position. The driver uses 1200 combs and the electrostatic force is 210μN. Akira et al. [87] designed a lateral movement shunt DC switch. The total dimension of this switch is 3×1.5×0.5mm³ (the length, width and thickness respectively). The switch has a frequency band from 0 to 75 GHz and the actuation voltage is 5V. The switching time is 10.3μs. Park et al. [88] proposed a lateral movement capacitive shunt RF MEMS switch for 23.5 to 29GHz. Flexible structure is a folded beam spring. The actuation voltage of this switch is 25V and the switching time is 8ms. It used 1000 combs with a

gap of 2.1μm. The air is used as both on and off state capacitive coupling switch instead of dielectric material.

3.2 Spring constant

According to Table 4, spring constant plays an important role on the actuation voltage of RF MEMS switches. The spring constant of MEMS switch is given by

$$K = K_{spring} + K_{\sigma} \qquad (8)$$

Therefore, reduction of the spring constant can be categorized into reduction of spring constant of beam or membrane and residual stress (K_{σ}).

3.2.1 Spring constant of beam or membrane

Beam or membrane spring constant consists of two parameters: K_{spring}, which is due to material properties such as Young's modulus and shapes of the beam or membrane. A comprehensive study of low constant beams such as fixed-fixed, crab leg or folded flexures that can be used for reducing the actuation voltage is provided in [71]. Kundu et al. [89] reported a low actuation voltage RF MEMS switch with a frequency band from 5GHz to 30GHz. They introduced the concept of moving transmission line and membrane. Therefore, the equivalent spring constant of the switch follows the series spring constant rules. The actuation voltage reduced from 20V to 15V (Figure 7).
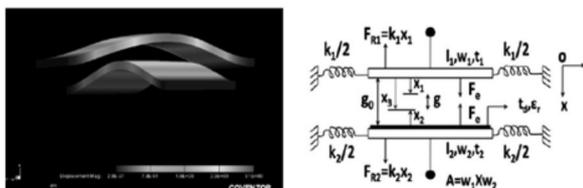
Figure 7: RF MEMS switch with movable electrode and mechanical model [89]

3.2.2 Reduction of residual stress

The second parameter is the effect of tensile residual stress on the spring constant during the fabrication process [90]. It causes the beam to deflect upward, therefore it increases the actuation voltage. The residual stress can be reduced by different techniques. The first method is the cancelling of the residual stress by using different micro structures. Zhiao et al. [91] reported a RF MEMS series switch for DC to X band which used AL/Au slant beam at the end of the cantilever. This composite beam is able to cancel the bending moment. The actuation voltage of this switch is 40V. Ur Rahman et al. [92] used a dimple at the end of the beam to overcome the residual stress. The contact of this switch is DC and it is fabricated on Alumina with a CPW transmission line for the frequencies of DC-40

GHz. The actuation voltage of the switch is 19V. Chan et al. [93] designed an inline low actuation voltage by reducing the sensitivity of the beam to residual stress through applying corrugations to their beam. This switch is a DC contact series RF MEMS switch with an actuation voltage of 20V. This switch is supported at two anchor points. It also has four springs which are connected at one end to an anchor and at the other end to the centre beam.

The second method to reduce the residual stress is through low-stress fabrication processes. Gong et al. (2009) developed a flat cantilever for 2-75GHz. They used an Al base sacrificial layer instead of a polymer sacrificial layer to reduce the stress gradient for the gold membrane. This is due to the coefficient of thermal expansion (CTE) between Al and Au (21 and 14ppm/K), respectively, rather than the typical polymer materials such as photoresist (>50ppm/K). The actuation voltage of this switch is still high and more than 40V. Biyikli et al. [94] reported a DC contact RF MEMS series switch with a frequency band from 0 to 25GHz. The gap between the transmission line and the beam is controlled by the amount of internal stress gradient. Tuning of the stress gradient depends on the decrease and increase of pressure for the bottom half, which results in a compressive stressed layer, and increase of the pressure for the top half layer for achieving the tensile stress. This leads to a compressive and tensile stress for the bottom and top layers. This experiment was done on different sizes of cantilever with the length (L=5-50μm) and width (W=2-40μm). The actuation voltage of all switches in this experiment is less than 20V.

3.3 Reduction of size (carbon switches)

Carbon switches can be fabricated in the range of nm, therefore, this type of switches is mostly named as NEMS switches. The rule for calculation of mechanical and RF parameters of carbon switches follows the rule of MEMS switches. The only difference between the MEMS and the NEMS switches is the role of Van der Waals force (Figure 2). A dynamic and mechanical study on the CNT NEMS switch is presented in [95-97]. The actuation voltage of the NEMS switches based on CNT and graphene is given as bellow:
For CNT:

$$V_{pull-in} = \sqrt{\left[\left[k(g-g_{eq}) - \frac{\pi C_6 \rho^2 \omega L}{6}\left(\frac{1}{g_{eq}^3} - \frac{1}{(g_{eq}+t)^3}\right)\right]x \frac{2g_{eq}^2}{\left(1+\frac{2g_{eq}}{\pi \omega}\right)\omega L\varepsilon_0}\right]} \qquad (9)$$

where g is the gap between the conductor and the ground, C_6 is a constant characterizing the interac-

tions between the two atoms, ρ is the volume density of graphite, which is taken to be $\rho = 1.14 \times 1029$ m–3, and $g_{eq} = 2/3g_0$.

For Graphene:

$$V_{pull-in} = \sqrt{\frac{8kg_0^3}{27\varepsilon \, WL} - \frac{A_h}{2\pi\varepsilon g_0}}$$

(10)

where A_h is the Hamaker constant (1.579eV), W and L are physical dimension of cantilever

The first term in the Equation (10) represents the contribution of the electrostatic force and the second term refers to the contribution of the Van der Waals force.

The first type of CNT NEMS switch is a dual fixed type switch. Kaul et al. [98] reported on a dual fixed capacitive RF NEMS switch based on SWCNT. The actuation voltage of the switch is less than 5V and the switching time is 2.8ns. The size of the SWCNT beam is 200nm long, and with a diameter of 2nm, and a gap of 20nm. The Young's modulus for this switch is 1Tpa. Acquaviva et al. [64] reported on a dual fixed capacitive RF NEMS switch based on SWCNT arrays for membrane. The actuation voltage of this switch is 6V. The resistivity of the beam is reported as 0.0077 Ω.cm and the flexural Young's modulus is very low (8.5 GPa). This is due to the fact that only a small portion of CNT contributes as a membrane and shear modulus during the actuation. A very low actuation voltage and fast dual fixed type RF NEMS switch is reported by Dragoman et al. [99]. The actuation voltage of this switch is less than 1 volt and its switching time is 100ps.

Another type of RF NEMS switch is the cantilever type RF NEMS switch. Dragoman et al. [99] reported on a RF cantilever NEMS switch which used 4 vertical CNT cantilevers based on the CPW as a nanotweezer switch. Each two cantilevers are attracted by applying a DC voltage across them and forming a short circuit. The length of CNT tweezers is 2.5μm. The actuation voltage for this switch is 14.5V, and is higher than other report-

ed CNT NEMS switches. This is due to the low Van der Waals force. As discussed before, Van der Waals force is only effective in nm gaps. The switching time for this switch is 49ns. Lee et al. [100] reported on a cantilever type RF NEMS switch based on MWCNT. The actuation voltage of this switch is less than 5V. The CNT has a 0.5nm diameter and is 1.8μm long. The gap between the transmission line and the CNT cantilever is 150nm. This switch used gold as a bottom electrode and Au/Ti (70/5nm) for CNT contact.

Milaninia et al. [101] presented the NEMS switch with two layers of graphene. Therefore, two CVD processes were used. The size of the beam is 20 × 3μm (L × w) and g = 500nm. The actuation voltage is 4.5V. The main disadvantage of this switch is a limitation of contact resistance between the top and bottom graphene layers (200kΩ). This is due to the nonuniform surface of the CVD grown graphene. Dragoman et al. [102, 103] simulated a double clamped RF NEMS switch based on a graphene membrane. 20nm gold is patterned on 500μm Si to form a CPW transmission line. The gap between the signal line and the membrane is 1μm. The CPW is loaded by the number of graphene flakes with a width of 0.6μm. This switch can be used for 1-60GHz applications. Increasing the number of graphene membranes above the TL increases the performance of the switch in the down state position, but it does not affect it on the up state position. The actuation voltage of this switch is 2V. The main drawback of all the discussed RF NEMS switches is the insertion loss and isolation in up and down state positions, respectively.

4 Discussion

As described in Section 3, the actuation voltage of the RF MEMS can be reduced by using three methods. Table 5 compares the impacts of these three methods on both mechanical and RF parameters of the switches.

Impacts on Actuation voltage
Table 5 compares the effect of each method on the actuation voltage of RF MEMS switches. As can be seen

Table 5: Comparison of low actuation methods on mechanical and RF parameters

Methods / Parameters		Vpull-in	RF parameters	Switching time	Fabrication process	Reduction of size
Reduction of Spring constant	Beam	Low	Good	Slow	Hard	Medium
	Residual stress	High	Good	Slow	Hard	Medium
Reduction of Gap	Matching circuit	Low	Very Good	Good	Good	Large
	Torsional actuation	Low	Good	Good	Good	Large
	Comb structure	Low	Good	Very Slow	Simple	Very large
Reduction of Size	Carbon Nano tube	Extremely Low	Not Good	Extremely fast	Extremely Hard	Small
	Graphene	Extremely Low	Not Good	Extremely Fast	Extremely hard	Small

from the table, carbon switches have the lowest actuation voltage. Decreasing the size has a negative impact on the actuation voltage but because of the presence of the Van der Waals force in the range of nm, this type of switches has the lowest actuation voltage. The second method that has the highest impact on lowering the actuation voltage is reduction of the gap because the relation between the gap and the actuation voltage is more than other parameters ($V_{pull-in} \propto g^{\frac{3}{2}}$). The last method is to reduce the spring constant. This method has the lowest impact on the actuation voltage ($V_{pull-in} \propto K^{\frac{1}{2}}$).

RF parameters

Table 5 compares the effects of each method on the RF parameters. The first parameter is reduction of the gap. The RF parameters of the switch strongly depend on the amount of capacitance in the up and down states. As presented in Table 4, the reduction of the gap has a negative impact on RF parameters. This problem is resolved by the techniques that were reviewed in Section 3. For example, compensation of the parasitic capacitance by matching impedance, providing a large contact area and a large gap by torsional actuations or increasing the force by lateral comb driver. These techniques are used to reduce the gap with acceptable RF parameters. Also, low actuation can be achieved by reducing the K without any effect on the amount of capacitance and RF parameters. The RF parameters of NEMS switches are not good for the microwave frequency but this may change in the future.

Switching time

Table 5 compares the effects of each method on switching time. NEMS switches based on CNT and Graphene have the highest speed due to the high Young's modulus of CNT and graphene, which is 1TPa. It is reported that the speed of the NEMS switches is in the range of a few tens of nano seconds and even equal to or faster than semiconductor switches [99, 104]. Using torsional and matching impedances does not have an effect on the speed of switches, and this is the same method used for other types of conventional RF MEMS switches. But comb structures are very slow because all actuators, transmission lines and beams are connected to each other, and therefore the membrane is very heavy compared to other type of MEMS switches. Reducing the spring constant also impacts negatively on the speed of the switch.

Size of switch

Table 5 compares the size of the RF MEMS switches. It is obvious that the size of RF NEMS switches is smaller than the MEMS switches. Lowering the spring constant does not affect the size of switches. However, methods for lowering the actuation voltage by decreasing the gap increase the size of MEMS switches. This is due to the micro structures that they use for reducing the actuation voltage. For example, a comb structure has the largest size because of the drivers. Moreover, reduction of the gap by using matching circuits uses additional space on the transmission line due to the presence of SHITL. Torsional actuation has a large area due to the pillars, extra electrodes and additional voltage source.

Fabrication and set up

The fabrication process of RF MEMS switches based on CNT and Graphene is different to that of conventional RF MEMS switches and is more sophisticated than that used for conventional MEMS switches. Also, CNT and graphene are more expensive than other metal materials.

The fabrication process for reduction of the gap is categorized into three methods. The first method for reduction of the gap uses matching impedance. The fabrication of this type exactly follows the fabrication of the RF MEMS switch and there is no additional process for it. The second group uses torsional actuation. It requires more fabrication processing than the normal process for MEMS switches. For example, the switch which is fabricated by Touati et al. [80] used nine masks and RIE etching for patterning pillars. Moreover, it requires two voltage sources to provide positive and negative voltages. The third type is comb switches. The fabrication process is less complex than other types of MEMS switches because actuator, transmission line and beam are all fabricated in one step of lithography. However, the gap demands RIE etching instead of wet etching. Moreover, most of the reported RF MEMS comb switches are DC contact and there is less capacitive MEMS shunt switches reported with comb structures. This is due to the deposition of dielectric on the side walls which limits the On and Off capacitance ratio. The existing capacitance RF MEMS comb switches were fabricated by Park et al. [88], where air was used as dielectric material. He et al. [85] used paryline material instead of air.

The methods for reduction of spring constant can be categorized in two groups. The first method uses different types of micro structures such as pillars or corrugation to cancel the curling. It requires an additional fabrication process. The second method is fabrication of MEMS switches via low residual stress material. The main disadvantage of this method is the complexity of measuring the amount of residual stress. This is because it requires determination of the exact amount of residual stress on each step such as lithography and deposition of materials.

5 Conclusion

In this paper, methods for the reduction of the actuation voltage of RF MEMS electrostatic switches have been studied. The study was conducted based on various experiments and analysis presented in recent published works. Electrostatic MEMS/NEMS switches are categorized based on their materials into metallic and carbon switches. Reduction of gap and spring constant are mostly used for reducing the actuation voltage of RF MEMS switches based on metals. The fabrication of this type of switches is based on surface micromachining. Switches based on CNT and graphene, NEMS switches, are fabricated in nano sizes. They are a new generation of electro-mechanical switches, and researchers are trying to improve their RF parameters. The fabrication of this type of switches is based on the CVD process. The impact of this method has been analysed and briefly discussed according to the mechanical parameters and RF parameters, taking into account the fabrication process. The materials presented in this paper enable researchers to better optimize their design based on the available fabrication facility and the desired application.

References

1. Lucyszyn, S., Advanced RF Mems, 2010. (New York: Cambridge University Press).
2. Rebeiz, G.M., RF MEMS, 2003. Wiley Online Library).
3. Gammel, P., G. Fischer and J. Bouchaud, RF MEMS and NEMS technology, devices, and applications, Bell Labs Technical Journal, 2005. 10. p. 29-59.
4. Rebeiz, G., K. Entesari, I. Reines, S.J. Park, M. El-Tanani, A. Grichener and A. Brown, Tuning in to RF MEMS, Microwave Magazine, IEEE, 2009. 10. p. 55-72.
5. Jones, R. and M. Chapman, RF MEMS in mobile phones, RF DESIGN, 2005. 28. p. 20.
6. Gu, Q. and J.R. De Luis. RF MEMS tunable capacitor applications in mobile phones. 2010. IEEE.
7. Kolb, S., MEMS PRODUCTS AND MEMS TECHNOLOGIES FOR AUTOMOTIVE APPLICATIONS AT INFENION, INFORMACIJ MIDEM, 2006. 36. p. 185-189.
8. Drago Strle, V.K., MEMS BASED INERTIAL SYSTEMS, INFORMACIJ MIDEM, 2007. 37. p. 11.
9. Mortazavi, D., A.Z. Kouzani, Y. Mafinejad and M. Hosain. Plasmon eignevalues as a function of nano-spheroids size and elongation. Complex Medical Engineering (CME), 2012 ICME International Conference on, 2012. IEEE.
10. Khoshmanesh, K., A.Z. Kouzani, S. Nahavandi, S. Baratchi and J. Kanwar, Design and simulation of an interdigital-chaotic advection micromixer for lab-on-a-chip applications, La houille blanche, 2009. 118-124.
11. Tehranirokh, M., A.Z. Kouzani, P.S. Francis and J.R. Kanwar, Generating different profiles of gradient concentrations inside a gel-filled chamber: design and simulation, Microsystem technologies, 2012. 1-6.
12. Islam, M., A.Z. Kouzani, X.J. Dai, W.P. Michalski and H. Gholamhosseini, Design and Analysis of a Multilayer Localized Surface Plasmon Resonance Graphene Biosensor, Journal of Biomedical Nanotechnology, 2012. 8. p. 380-393.
13. Daneshmand, M. and R. Mansour, RF MEMS Satellite Switch Matrices, Microwave Magazine, IEEE, 2011. 12. p. 92-109.
14. Malmqvist, R., C. Samuelsson, B. Carlegrim, P. Rantakari, T. Vähä-Heikkilä, A. Rydberg and J. Varis. Ka-band RF MEMS phase shifters for energy starved millimetre-wave radar sensors. 2010. IEEE.
15. Fox, C.H., X. Chen, H.W. Jiang, P.B. Kirby and S. McWilliam. Development of micromachined RF switches with piezofilm actuation. 2002.
16. Polcawich, R.G., J.S. Pulskamp, D. Judy, P. Ranade, S. Trolier-McKinstry and M. Dubey, Surface micromachined microelectromechancial ohmic series switch using thin-film piezoelectric actuators, Microwave Theory and Techniques, IEEE Transactions on, 2007. 55. p. 2642-2654.
17. Gross, S., S. Tadigadapa, T. Jackson, S. Trolier-McKinstry and Q. Zhang, Lead-zirconate-titanate-based piezoelectric micromachined switch, Applied Physics Letters, 2003. 83. p. 174-176.
18. Guerre, R., U. Drechsler, D. Bhattacharyya, P. Rantakari, R. Stutz, R.V. Wright, Z.D. Milosavljevic, T. Vaha-Heikkila, P.B. Kirby and M. Despont, Wafer-level transfer technologies for PZT-based RF MEMS switches, Journal of Microelectromechanical Systems, 2010. 19. p. 548-560.
19. Proie, R.M., R.G. Polcawich, J.S. Pulskamp, T. Ivanov and M.E. Zaghloul, Development of a PZT MEMS Switch Architecture for Low-Power Digital Applications, Microelectromechanical Systems, Journal of, 2011. 20. p. 1032-1042.
20. Polcawich, R.G., D. Judy, J.S. Pulskamp, S. Trolier-McKinstry and M. Dubey. Advances in Piezoelectrically Actuated RF MEMS Switches and Phase Shifters. Microwave Symposium, 2007. IEEE/MTT-S International, 2007.
21. Mahameed, R., N. Sinha, M.B. Pisani and G. Piazza, Dual-beam actuation of piezoelectric AlN RF MEMS switches monolithically integrated with AlN contour-mode resonators, Journal of Micro-

mechanics and Microengineering, 2008. 18. p. 105011.

22. Klaasse, G., R. Puers and H. Tilmans, Piezoelectric versus electrostatic actuation for a capacitive RF-MEMS switch, Proc. SeSens, 2002. 631-634.

23. Lee, T.M., Y.H. Seo, K.H. Whang and D.S. Choi, Study on the Lateral Piezoelectric Actuator with Actuation Range Amplifying Structure, Key Engineering Materials, 2006. 326. p. 289-292.

24. Il-Joo, C., S. Taeksang, B. Sang-Hyun and Y. Euisik, A low-voltage and low-power RF MEMS series and shunt switches actuated by combination of electromagnetic and electrostatic forces, Microwave Theory and Techniques, IEEE Transactions on, 2005. 53. p. 2450-2457.

25. Cho, I.J. and E. Yoon, Design and fabrication of a single membrane push-pull SPDT RF MEMS switch operated by electromagnetic actuation and electrostatic hold, Journal of Micromechanics and Microengineering, 2010. 20. p. 035028.

26. Zhang, Y., G. Ding, X. Shun, D. Gu, B. Cai and Z. Lai, Preparing of a high speed bistable electromagnetic RF MEMS switch, Sensors and Actuators A: Physical, 2007. 134. p. 532-537.

27. Nordquist, C.D., M.S. Baker, G.M. Kraus, D.A. Czaplewski and G.A. Patrizi, Poly-Silicon Based Latching RF MEMS Switch, IEEE Microwave and Wireless Components Letters, 2009. 19. p. 380-382.

28. de los Santos, H.J., G. Fischer, H.A.C. Tilmans and J.T.M. Van Beek, RF MEMS for ubiquitous wireless connectivity. Part I. Fabrication, Microwave Magazine, IEEE, 2004. 5. p. 36-49.

29. Balachandran, S., J. Kusterer, D. Maier, M. Dipalo, T. Weller and E. Kohn. High power nanocrystalline diamond RF MEMS- A combined look at mechanical and microwave properties. Microwaves, Communications, Antennas and Electronic Systems, 2008. COMCAS 2008. IEEE International Conference on, 2008.

30. Seong-Dae, L., J. Byoung-Chul, S.D. Kim and R. Jin-Koo, A novel pull-up type RF MEMS switch with low actuation voltage, Microwave and Wireless Components Letters, IEEE, 2005. 15. p. 856-858.

31. Mansour, R., M. Bakri-Kassem, M. Daneshmand and N. Messiha. RF MEMS devices. 2003. IEEE.

32. Ruan, J., G.J. Papaioannou, N. Nolhier, M. Bafleur, F. Coccetti and R. Plana. ESD stress in RF-MEMS capacitive switches: The influence of dielectric material deposition method. Reliability Physics Symposium, 2009 IEEE International, 2009.

33. Malmqvist, R., C. Samuelsson, W. Simon, P. Rantakari, D. Smith, M. Lahdes, M. Lahti, Va, x, ha, Heikkila, T., J. Varis and R. Baggen. Design, packaging and reliability aspects of RF MEMS circuits fabricated using a GaAs MMIC foundry process tech-nology. Microwave Conference (EuMC), 2010 European, 2010.

34. Das, C.K., P. Bhattacharya and S.S. Kalra, Graphene and MWCNT: Potential Candidate for Microwave Absorbing Materials, Journal of Materials Science Research, 2012. 1. p. p126.

35. Dragoman, M., D. Neculoiu, D. Dragoman, G. Deligeorgis, G. Konstantinidis, A. Cismaru, F. Coccetti and R. Plana, Graphene for Microwaves, Microwave Magazine, IEEE, 2010. 11. p. 81-86.

36. Liao, M. and Y. Koide, Carbon-Based Materials: Growth, Properties, MEMS/NEMS Technologies, and MEM/NEM Switches, Critical Reviews in Solid State and Materials Sciences, 2011. 36. p. 66-101.

37. Saberi, M., R. Lotfi, K. Mafinezhad and W.A. Serdijn, Analysis of power consumption and linearity in capacitive digital-to-analog converters used in successive approximation ADCs, Circuits and Systems I: Regular Papers, IEEE Transactions on, 2011. 58. p. 1736-1748.

38. Saberi, M., H. Sepehrian, R. Lotfi and K. Mafinezhad, A low-power Successive Approximation ADC for biomedical applications, IEICE electronics express, 2011. 8. p. 195-201.

39. Mafinezhad, K., Modeling and optimisation of a solenoidal integrated inductor for RFICs, International journal of RF and Microwave Computer-Aided Engineering, 2009. 5. p.

40. Mafinezhad, K. and S.H. Keshmiri, Design and Simulation of an Oblique Suspender MEMS Variable Capacitor, Scientia Iranica, 2006. 13. p.

41. Nabovati, H., K. Mafinezhad, A. Nabovati and H. Keshmiri, Comprehensive Electromechanical Analysis of MEMS Variable Gap Capacitors, JOURNAL OF IRANIAN ASSOCIATION OF ELECTRICAL AND ELECTRONICS ENGINEERS, 2007. 4. p. 3.

42. Daliri, M., M. Maymandi-Nejad and K. Mafinezhad, Distortion analysis of bootstrap switch using volterra series, Circuits, Devices & Systems, IET, 2009. 3. p. 359-364.

43. Fedder, G.K., R.T. Howe, T.J.K. Liu and E.P. Quévy, Technologies for cofabricating MEMS and electronics, Proceedings of the IEEE, 2008. 96. p. 306-322.

44. Kris Baert, C.V.H., Integrated Microsystems, Informacij MIDEM, 2006. 36. p. 7.

45. Becker, J., Configurability for Systems on Silicon: Requirement and Perspective for future VLSI Solutions, INFORMACIJE MIDEM-LJUBLJANA-. 2003. 33. p. 236-244.

46. Lee, M.J., Y. Zhang, C. Jung, M. Bachman, F. De Flaviis and G. Li, A novel membrane process for RF MEMS switches, Microelectromechanical Systems, Journal of, 2010. 19. p. 715-717.

47. Goldsmith, C.L., Z. Yao, S. Eshelman and D. Denniston, Performance of low-loss RF MEMS capacitive

switches, Microwave and Guided Wave Letters, IEEE, 1998. 8. p. 269-271.

48. Lakshminarayanan, B. and T.M. Weller, Design and modeling of 4-bit slow-wave MEMS phase shifters, Microwave Theory and Techniques, IEEE Transactions on, 2006. 54. p. 120-127.

49. Mercier, D., K. Van Caekenberghe and G.M. Rebeiz. Miniature RF MEMS switched capacitors. 2005. IEEE.

50. G.Nassiopoulou, A., POROUS SILICON FOR SENSORS AND ON-CHIP INTEGRATION OF RF COMPONENT, Informacij MIDEM, 2006. 36. p. 7.

51. Ding, Y., Z. Liu, L. Liu and Z. Li, A surface micromachining process for suspended RF-MEMS applications using porous silicon, Microsystem technologies, 2003. 9. p. 470-473.

52. Guo, F., Z. Zhu, Y. Long, W. Wang, S. Zhu, Z. Lai, N. Li, G. Yang and W. Lu, Study on low voltage actuated MEMS rf capacitive switches, Sensors and Actuators A: Physical, 2003. 108. p. 128-133.

53. Meyyappan, M., Carbon nanotubes: science and applications, 2005. CRC).

54. Dresselhaus, M.S., G. Dresselhaus and P. Eklund, Science of fullerenes and carbon nanotubes, 1996. Academic Pr).

55. Dubois, S.M.M., Z. Zanolli, X. Declerck and J.C. Charlier, Electronic properties and quantum transport in Graphene-based nanostructures, The European Physical Journal B-Condensed Matter and Complex Systems, 2009. 72. p. 1-24.

56. Bolotin, K.I., K. Sikes, Z. Jiang, M. Klima, G. Fudenberg, J. Hone, P. Kim and H. Stormer, Ultrahigh electron mobility in suspended graphene, Solid State Communications, 2008. 146. p. 351-355.

57. Teo, K., M. Chhowalla, G. Amaratunga, W. Milne, D. Hasko, G. Pirio, P. Legagneux, F. Wyczisk and D. Pribat, Uniform patterned growth of carbon nanotubes without surface carbon, Applied Physics Letters, 2001. 79. p. 1534.

58. Gad-el-Hak, M., MEMS: design and fabrication, 2006, 2. CRC press).

59. Rahman, H.U., Plasma Based Dry Release of MEMS Devices.

60. Yo-Tak, S., L. Hai-Young and M. Esashi, Low actuation voltage capacitive shunt RF-MEMS switch having a corrugated bridge, IEICE transactions on electronics, 2006. 89. p. 1880-1887.

61. Giacomozzi, F., V. Mulloni, S. Colpo, J. Iannacci, B. Margesin and A. Faes, A Flexible Fabrication Process for RF MEMS Devices, Romanian Journal of Information Science and Technology (ROMJIST), 2011. 14. p. 259-268.

62. Iliescu, C., MICROFLUIDICS IN GLASS: TECHNOLOGIES AND APPLICATIONS, Informacij MIDEM, 2006. 36. p. 7.

63. Alireza Bahadorimehr, B.Y.M., FABRICATION OF GLASS-BASED MICROFLUIDIC DEVICES WITH PHOTORESIST AS MASK, INformacij MIDEM, 2011. 41. p. 4.

64. Acquaviva, D., A. Arun, S. Esconjauregui, D. Bouvet, J. Robertson, R. Smajda, A. Magrez, L. Forro and A. Ionescu, Capacitive nanoelectromechanical switch based on suspended carbon nanotube array, Applied Physics Letters, 2010. 97. p. 233508.

65. Cassell, A.M., N.R. Franklin, T.W. Tombler, E.M. Chan, J. Han and H. Dai, Directed growth of free-standingsingle-walled carbon nanotubes, Journal of the American Chemical Society, 1999. 121. p. 7975-7976.

66. Mahshid Kalani, R.Y., carbon nano tube via chemical vapour deposition, Asian Journal of Chemistry, 2011. 23. p. 4735-4743.

67. Dequesnes, M., S. Rotkin and N. Aluru, Calculation of pull-in voltages for carbon-nanotube-based nanoelectromechanical switches, Nanotechnology, 2002. 13. p. 120.

68. Dequesnes, M., Z. Tang and N. Aluru, Static and dynamic analysis of carbon nanotube-based switches, Journal of engineering materials and technology, 2004. 126. p. 230.

69. Kang, J.W., S.C. Kong and H.J. Hwang, Electromechanical analysis of suspended carbon nanotubes for memory applications, Nanotechnology, 2006. 17. p. 2127.

70. Lucyszyn, S., Advanced RF MEMS 2010. Cambridge University Press).

71. Rebeiz, G.M., RF MEMS: Theory, Design, and Technology, 2003.

72. Varadan, V.K., RF MEMS and their applications 2002, Wiley: New York.

73. Younis, M.I., Sensing and Actuation in MEMS, MEMS Linear and Nonlinear Statics and Dynamics, 2011. 57-96.

74. Reinhold Ludwig, G.B., RF circuit design : theory and applications 2009. Prentice-Hall).

75. Mafinejad, Y., A.Z. Kouzani, K. Mafinezhad and A. Golmakani, Pi-shaped MEMS architecture for lowering actuation voltage of RF switching, IEICE electronics express, 2009. 6. p. 1483-1489.

76. Mafinejad, Y., K. Mafinezhad and A.Z. Kouzani, Improving RF characteristics of MEMS capacitive shunt switches, International Review of Modelling and Simulations (IREMOS), 2009.

77. Mafinejad, Y., A.Z. Kouzani, K. Mafinezhad and H. Nabovatti. Design and simulation of a low voltage wide band RF MEMS switch. International Conference on Systems, Man and Cybernetics San Antonio, Texas, USA, 2009. IEEE.

78. Mafinejad, Y., A.Z. Kouzani, K. Mafinezhad and A. Kaynak, Low Actuation Wideband RF MEMS

Shunt Capacitive Switch, Procedia Engineering, 2012. 29. p. 1292-1297.

79. Zarghami, M., Y. Mafinejad, A. Kouzani and K. Mafinezhad, Low actuation-voltage shift in MEMS switch using ramp dual-pulse, IEICE electronics express, 2012. 9. p. 1062-1068.

80. Touati, S., N. Lorphelin, A. Kanciurzewski, R. Robin, A.S. Rollier, O. Millet and K. Segueni. Low actuation voltage totally free flexible RF MEMS switch with antistiction system. 2008. IEEE.

81. Robin, R., S. Touati, K. Segueni, O. Millet and L. Buchaillot. A new four states high deflection low actuation voltage electrostatic MEMS switch for RF applications. 2008. IEEE.

82. Kim, C., Mechanically Coupled Low Voltage Electrostatic Resistive RF Multi-throw Switch, IEEE Transactions on Industrial Electronics, 2011. 1-1.

83. Liu, A., M. Tang, A. Agarwal and A. Alphones, Low-loss lateral micromachined switches for high frequency applications, Journal of Micromechanics and Microengineering, 2005. 15. p. 157.

84. He, X., B. Liu, Z. Lv and Z. Li, A lateral RF MEMS capacitive switch utilizing parylene as dielectric, Microsystem Technologies, 2012. 1-9.

85. He, X., B. Liu, Z. Lv and Z. Li, A lateral RF MEMS capacitive switch utilizing parylene as dielectric, Microsystem Technologies, 2011. 1-9.

86. Kang, S., H.C. Kim and K. Chun. Single pole four throw RF MEMS switch with double stop comb drive. 2008. IEEE.

87. Akiba, A., S. Mitarai, S. Morita, K. Ikeda, S. Kurth, S. Leidich, A. Bertz, M. Nowack, J. Froemel and T. Gessner. A fast and low actuation voltage MEMS switch for mm-wave and its integration. 2010. IEEE.

88. Park, J., E.S. Shim, W. Choi, Y. Kim, Y. Kwon and D. Cho, A Non-Contact-Type RF MEMS switch for 24-GHz radar applications, Journal of Microelectromechanical Systems, 2009. 18. p. 163-173.

89. Kundu, A., S. Sethi, N. Mondal, B. Gupta, S. Lahiri and H. Saha, Analysis and optimization of two movable plates RF MEMS switch for simultaneous improvement in actuation voltage and switching time, Microelectronics Journal, 2010. 41. p. 257-265.

90. Huang, J.M., K.M. Liew, C.H. Wong, S. Rajendran, M.J. Tan and A.Q. Liu, Mechanical design and optimization of capacitive micromachined switch, Sensors and Actuators A: Physical, 2001. 93. p. 273-285.

91. Zhihao, H., L. Zewen and L. Zhijian. Al/Au composite membrane bridge DC-contact series RF MEMS switch. Solid-State and Integrated-Circuit Technology, 2008. ICSICT 2008. 9th International Conference on, 2008.

92. Rahman, H.U. and R. Ramer. Supported bars novel cantilever beam design for RF MEMS series switches. 2009. IEEE.

93. Chan, K.Y. and R. Ramer. RF MEMS Switch with low stress sensitivity and low actuation voltage. 2009. IEEE.

94. Biyikli, N., Y. Damgaci and B. Cetiner, Low-voltage small-size double-arm MEMS actuator, Electronics Letters, 2009. 45. p. 354-356.

95. Shi, Z., H. Lu, L. Zhang, R. Yang, Y. Wang, D. Liu, H. Guo, D. Shi, H. Gao and E. Wang, Studies of graphene-based nanoelectromechanical switches, Nano Research, 2011. 1-6.

96. Fujita, S., K. Nomura, K. Abe and T.H. Lee, 3-d nanoarchitectures with carbon nanotube mechanical switches for future on-chip network beyond cmos architecture, Circuits and Systems I: Regular Papers, IEEE Transactions on, 2007. 54. p. 2472-2479.

97. Rajter, R., R. French, W. Ching, W. Carter and Y. Chiang, Calculating van der Waals-London dispersion spectra and Hamaker coefficients of carbon nanotubes in water from ab initio optical properties, Journal of Applied Physics, 2007. 101. p. 054303.

98. Kaul, A.B., E.W. Wong, L. Epp and B.D. Hunt, Electromechanical carbon nanotube switches for high-frequency applications, Nano letters, 2006. 6. p. 942-947.

99. Dragoman, M., A. Takacs, A. Muller, H. Hartnagel, R. Plana, K. Grenier and D. Dubuc, Nanoelectromechanical switches based on carbon nanotubes for microwave and millimeter waves, Applied Physics Letters, 2007. 90. p. 113102-113102-3.

100. Lee, S.W., D.S. Lee, R.E. Morjan, S.H. Jhang, M. Sveningsson, O. Nerushev, Y.W. Park and E.E.B. Campbell, A three-terminal carbon nanorelay, Nano letters, 2004. 4. p. 2027-2030.

101. Milaninia, K.M., M.A. Baldo, A. Reina and J. Kong, All graphene electromechanical switch fabricated by chemical vapor deposition, Applied Physics Letters, 2009. 95. p. 183105.

102. Dragoman, M., D. Dragoman, F. Coccetti, R. Plana and A. Muller, Microwave switches based on graphene, Journal of Applied Physics, 2009. 105. p. 054309-054309-3.

103. Dragoman, M., D. Dragoman and A. Muller. High frequency devices based on graphene. 2007. IEEE.

104. Frank, I., D. Tanenbaum, A. Van Der Zande and P. McEuen, Mechanical properties of suspended graphene sheets, Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures, 2007. 25. p. 2558.

Cylindrical-parabolic reflector with printed antenna structures

Aleksandar Nešić[1], Ivana Radnović[1], Marija Milijic[2], Zoran Mićić[1] and Dusan Nešić[3]

[1]Institute IMTEL Komunikacije a.d, New Belgrade, Serbia
[2]Faculty of Electronic Engineering, University of Nis, Nis, Serbia

[3]IHTM-CMTM, University of Belgrade, Belgrade, Serbia

Abstract: The paper presents concept of design and realization of the new class of printed antenna structures which consist of a linear axial array of dipoles, subreflector, feed network and a bal-un, all printed on a common dielectric substrate. The array is positioned on the axis focus of the cylindrical-parabolic reflector. Use of the reflector enables reducing back side radiation and shaping beamwidth in H-plane thus obtaining higher gain while the printed subreflector gives the possibility of achieving additional gain. Besides, by using dipoles with pentagonal shape that operate on the second resonance, enhanced bandwidth of the array has been accomplished. Four variants of such arrays have been realized: two of them with 8 radiating elements for the frequency range around 26 GHz – one with uniform and the other with tapered feed distribution, featuring gains of 27.5 dBi and 25.7 dBi, respectively. The latter has the side lobe suppression of 28 dB in E-plane. Two other arrays that are intended for ranges around 23 GHz and 60 GHz have 16 radiating elements, uniform feed distribution and measured gains of 33 dBi and 34 dBi, respectively. Bandwidths of all realized model for S11 less than -10 dB is around 30 %. In all cases agreement between simulated and measured results is very good.

Key words: Microwaves and millimeter waves, Antenna array, Printed antenna, Cylindrical-parabolic reflector

Cilindrično parabolični odbojniki s tiskanimi antenam

Povzetek: Članek predstavlja koncept dizajna in realizacijo nove tiskane antene, ki je sestavljena z linearno matriko dipolov, odbojnikom in napajalnim omrežjem na skupnem dielektričnem substratu. Niz je nameščen v žariščni osi cilindričnega paraboličnega odbojnika. Uporaba odbojnika omogoča znižanje sevanja nazaj in oblikovanje pasovne širine v ravnini H za doseganje večjega ojačenja Uporaba dipolov v obliki peterokotnika in delovanju v drugi resonančni frekvenci omogoča povečanje pasovne širine. Realizirane so štiri inačice teh matrik. V vseh primerih je bilo doseženo dobro ujemanje med simulacijami in meritvami.

Ključne besede: mikrovalovi in milimeterski valovi, matrična antena, tiskana antena cilindričen paraboličen odbojnik

* Corresponding Author's e-mail: nesicad@nanosys.ihtm.bg.ac.rs

1 Introduction

In the last two decades printed antenna structures are dominant over conventional antennas especially in microwave and millimeter wave ranges, except in some specific applications. They are practically indispensable in the field of reconfigurable antennas. Main advantages of printed antenna structures are: high reproducibility, small dimensions, low weight, compactness, beam scanning possibility as well as possibility of integration with passive and active microwave circuits.

However, there are applications where printed antennas show certain disadvantages such as: applications in high power transmitters (possibility of dielectric breakdown), applications in high gain antennas (due to relatively high losses in printed feed lines) and in antennas with very high side lobe suppression (over 40 dB) due to critical dimension tolerances.

In order to overcome these significant disadvantages, the concept of linear axial antenna array positioned on a focal axis of a cylindrical-parabolic reflector is introduced [1-4].

As radiating elements in the antenna array printed dipoles of pentagonal shape operating on the second resonance are used. In this way a much wider band-

width than with conventional printed antenna arrays is obtained.

The presence of the cylindrical-parabolic reflector introduces a third dimension in a standard planar structure, and this is practically the only drawback comparing to conventional printed antenna arrays. However, by applying a subreflector which consists of two strips printed on both sides of the dielectric substrate, the depth of the cylindrical-parabolic reflector can be reduced and so the overall size of the antenna structure decreased.

Differently from Cassegrain antennas where subreflector covers the central part of the antenna aperture resulting in lower efficiency and side lobe suppression, the effect of aperture blockage is negligible in presented concept. Moreover, owing to the subreflector, more suitable illumination distribution, i.e. higher gain and better side lobe suppression in H-plane can be achieved.

2 Concept

2.1 Radiating Elements

Radiating elements in the array are printed pentagonally shaped dipoles. One half of dipoles is printed on one side of the dielectric substrate and their other half on the opposite substrate side, Fig.1, [5]. These dipoles operate on the second resonance (antiresonance) enabling much slower impedance variation with frequency than in case of operation on the first resonance. The dipole impedance can range from about 70 Ω to a few hundred ohms. Variations of real and imaginary parts of the pentagonal dipole impedance with frequency are shown in Fig. 2. The dipole is optimized to impedance around (100 + j0) Ω at the central frequency of 26 GHz. Relative dielectric constant (ε_r) and thickness (h) of the substrate the dipole is printed on are 2.1 and 0.254 mm (CuFlon), respectively. Impedance variation
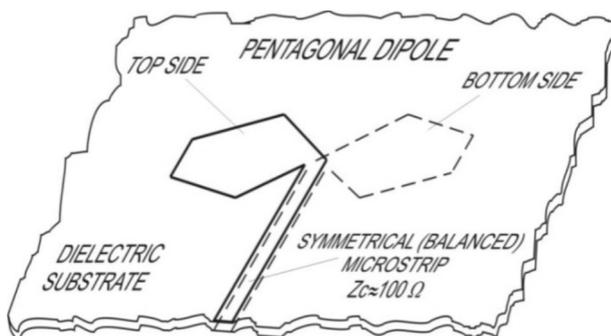
Figure 1: Printed pentagonal dipole as a basic element of the antenna array.

with frequency in such dipoles is a few dozen times smaller than in patches which are the most common conventional printed radiating elements. Bandwidth of the single dipole (VSWR < 2) is around 30 %.
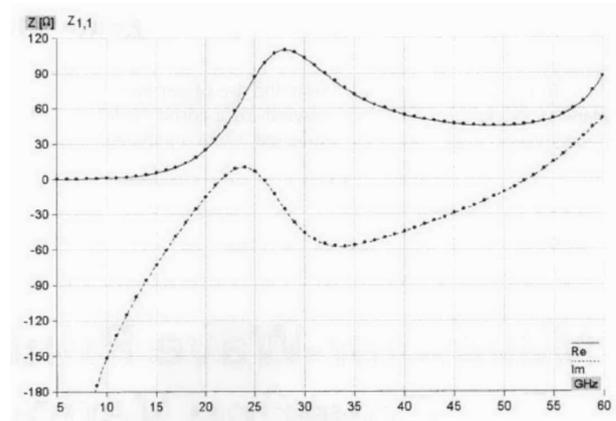
Figure 2: Real and imaginary parts of the pentagonal dipole impedance* as a function of frequency (ε_r=2.1, h =0.254 mm, S**= λ/4).

*The dipole impedance is optimized to obtain (100+j0) Ω at 26 GHz.
**S – distance between dipole axis and the flat reflector.

2.2 Axial Array of Dipoles

An axial array of dipoles has been formed, Fig. 3. Mutual impedances of these dipoles are relatively small which makes the design of the array significantly simpler which is especially important in cases of beam scanning. Certainly, the number of radiating elements in the array depends on the required E-field beamwidth. Distance between adjacent dipoles is practically a compromise between the gain and the sufficient suppression of grating lobes.

The distance between dipoles in axial antenna arrays with tapered feed distribution is chosen to be around 0.85 λ in order to obtain relatively high array gain and grating lobe level that is lower than the highest (usually the first) side lobe level, [6] (which is 13 dB suppressed with regard to the main lobe in case of uniform distribution).

2.3 Printed Subreflector

In front of the array of dipoles positioned on the focal axis of the cylindrical-parabolic reflector, there are strips printed on both sides of the substrate playing the role of a subreflector, Fig. 3, Fig. 5. Subreflector's axis and longitudinal axis of the dipoles' array are spaced λ/4 apart at the central frequency. Use of the subre-

flector offers the possibility of decreasing the depth of the cylindrical-parabolic reflector, i.e. of lowering the ratio L_f/D, where L_f is the focal length of the reflector and D is its width. Aperture blockage of the antenna structure by this subreflector is almost negligible due to extremely small thickness of the dielectric substrate (around $\lambda/100$). By varying width of the subreflector strip one can optimize illumination distribution of the cylindrical-parabolic reflector and obtain higher gain.

Figure 3: Printed antenna array with the subreflector and the feed network integrated on the same dielectric substrate.

2.4 Transition to Coaxial Connector or to Rectangular Waveguide

Complete feed network is realized in symmetrical (balanced) microstrip lines. In order to link a symmetrical to an asymmetrical (conventio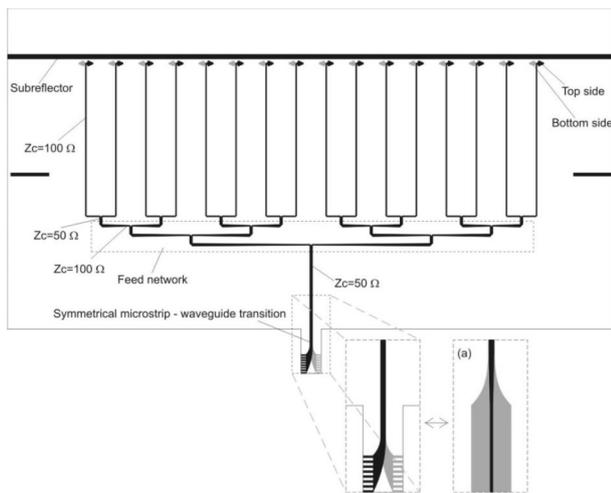nal) structure, a tapered symmetrical to asymmetrical microstrip transition (BAL-UN) is used in ranges below 40 GHz, Fig. 3(a). Its asymmetrical end is terminated with a coaxial SMA connector. However, in frequency ranges above 40 GHz there is a need for symmetrical microstrip-to-rectangular waveguide transition, [7]. Design of this transition is based on gradually tapering lines of a symmetrical microstrip that enters into waveguide and forming ridges at opposite sides of a dielectric substrate, Fig. 3. Through these ridges the E-field vector is being concentrated and rotated by 90° becoming parallel to shorter sides of a rectangular waveguide.

2.5 Cylindrical-Parabolic Reflektor

Cylindrical-parabolic reflector is made of a relatively thin aluminum (2 mm). It consists of two halves that are attached to each other along their apexes and fastened by screws or clasps. Holes through which symmetrical microstrip lines of the feed network pass are drilled through the junction of the two reflector's halves. Diameter (d) of these holes is chosen to be d>3W, where W is width of the symmetrical microstrip feed line, Fig 4. In this way the influence of the holes' rims on the feed line is being eliminated.

Ratio L_f/D (L_f – focal length of the reflector, D – width of the reflector) equals 0.2 in realized antennas while the length of the reflector (L) depends on the length of the axial array.

Figure 4: Holes in the cylindrical-parabolic reflector through which microstrip lines of the feed network pass.

2.6 Feed Network

Feed network is realized in symmetrical microstrip mainly for these reasons:
- Radiating elements are dipoles whose halves are printed on the opposite sides of the dielectric substrate thus representing a typical symmetrical structure.
- Feed networks in printed antenna structures are the main source of losses due to loss in feed lines. Symmetrical microstrip structure by its nature has lower losses than a conventional asymmetrical microstrip.

In realized antenna models with uniform feed distribution tapered impedance transformers are used in branching lines of the feed network, as shown in Fig 3. In case of antennas with high side lobe suppression we applied quarter-wave impedance transformers with suitable transformation ratio enabling desired tapered distribution, Fig 5.
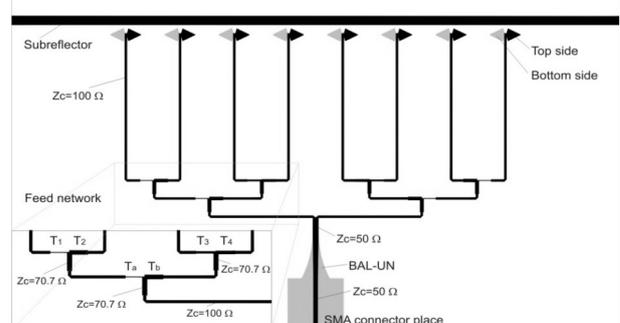

Figure 5: Example of feed network with impedance transformers (T_{1-4}, T_{a-b}) for tapered distribution (detail in the left lower corner).

Symmetrical microstrip feed networks are terminated either with symmetrical-to-asymmetrical microstrip transition or with symmetrical microstrip-to-waveguide transition, both described in the subchapter 2.4.

3 Realizations

On the basis of new design concept and by use of WIPL-D software package [8] for simulations four models of printed antenna arrays in the cylindrical-parabolic reflector have been developed:

(A) Array with 8 radiating elements operating at 26 GHz range with uniform feed distribution,

(B) Array with 8 radiating elements operating at 26 GHz range with tapered feed distribution,

(C) Array with 16 radiating elements operating at 23 GHz range with uniform feed distribution, and

(D) Array with 16 radiating elements operating at 60 GHz range with uniform feed distribution.

Their main technical and measured electrical characteristics are given in Table 1 while the photographs of the realized models are shown in Figures 6 - 9.

Figure 6: Photograph of the realized 26 GHz antenna array (A) in the cylindrical-parabolic reflector with uniform feed distribution.

From the obtained results we can come to a conclusion that practically all relevant parameters of this class of antenna arrays (bandwidth, gain, aperture efficiency) are significantly better than those of conventional printed antenna arrays.

Table 1

		Array (A)	Array (B)	Array (C)	Array (D)
Number of radiating elements		8	8	16	16
Feed distribution		Uniform	Tapered	Uniform	Uniform
Dielectric substrate		CuFlon (εr=2.1, h=0.254 mm, tanδ=4x10-4)	CuFlon (εr=2.1, h=0.254 mm, tanδ=4x10-4)	CuFlon (εr=2.1, h=0.254 mm, tanδ=4x10-4)	CuFlon (εr=2.1, h=0.127 mm, tanδ=4x10-4)
Dipoles' impedances		~(100+j0) Ω	~(100+j0) Ω	~(100+j0) Ω	~(100+j0) Ω
Distance between dipoles		11 mm (0.95 λ, 26 GHz)	10 mm (0.85 λ, 26 GHz)	14 mm (1.07 λ, 23 GHz)	5.5 mm (1.1 λ, 60 GHz)
Width of the subreflector		1.7 mm	1.7 mm	3.1 mm	0.866 mm
Dimensions of the cylindrical-parabolic reflector	L (Length)	110 mm	110 mm	250 mm	100 mm
	D (Width)	100 mm	100 mm	200 mm	100 mm
Length of the BAL-UN/sym. mstrip to waveguide transition (case D)		8 mm	8 mm	8 mm	5.09 mm
Type of connector		SMA	SMA	SMA	waveguide WR-15
Measured gain at the central frequency		~27.5 dBi (@26 GHz)	~25.7 dBi (@26 GHz)	~33 dBi (@23 GHz)	~34 dBi (@60 GHz)
FSLSE* (measured)	FSLSH* (measured)	~13 dB	~20 dB	~28 dB	~22 dB	~13 dB	~14 dB	~13 dB	~17 dB
Bandwidth (VSWR<2)		(24.5-28.0) GHz	(24.2-29.2) GHz	(18-28) GHz	(57.5-75.0) GHz
Aperture efficiency		54.1%	35.8%	54.2%	54.2%
Losses in metallization and dielectric		~1.1 dB	~1.1 dB	~2.2 dB	near 3 dB
maximum handling power, theory/real		~67 W / 33 W	~67 W / 33 W	~67 W / 33 W	~20 W / 10 W

$FSLS_E{}^*$, $FSLS_H{}^*$ - First Side Lobe Suppression in E- and H-plane

Maximum working temperature for CuFlon is 175^{0}C, 150^{0}C above normal temperature of 25^{0}C. Substrate CuFlon is a soft substrate with low thermal conductivity, 0.25 W/m/^{0}C, comparing to high thermal conductivity silicon [11].
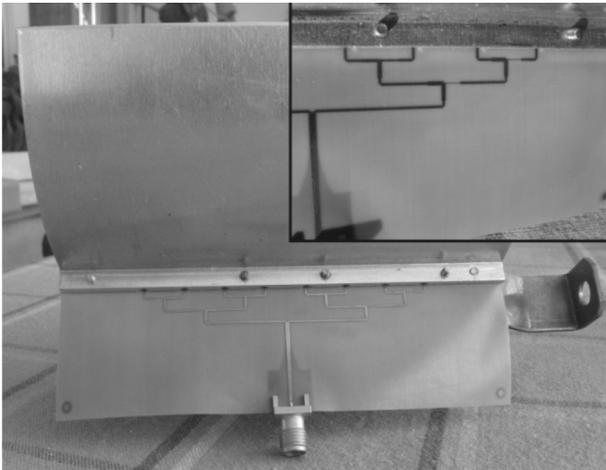
Figure 7: Photograph of the realized 26 GHz antenna array (B) in the cylindrical-parabolic reflector with tapered feed distribution with the detail of the tapered feed network.

Figure 8: Photograph of the realized 23 GHz antenna array (C) in the cylindrical-parabolic reflector with uniform feed distribution.
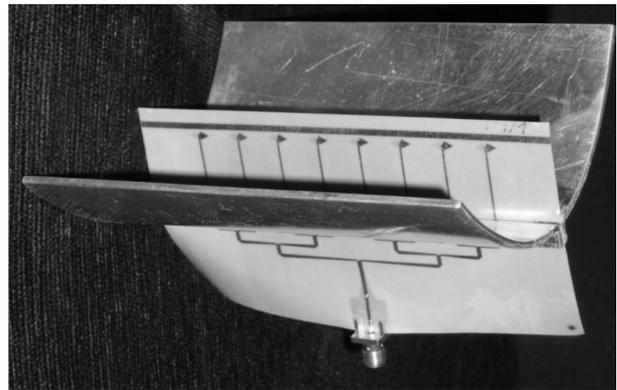

Figure 9: Photograph of the realized 60 GHz antenna array (D) in the cylindrical-parabolic reflector with uniform feed distribution.

4 Conclusion

The paper proposes a new class of printed antenna structures that have most of advantages of printed antennas: high reproducibility, low weight, compactness, possibility of simple integration with other passive and active microwave circuits as well as low manufacturing cost. At the same time they feature high gain and relatively low losses which is not common for planar printed antenna arrays. The only disadvantage comparing to standard planar (2D) antenna structures is the third dimension, i.e. thickness, due to presence of the cylindrical-parabolic reflector. Four realizations of printed antenna arrays in cylindrical-parabolic reflector intended for various frequency ranges are presented, including antenna with high side lobe suppression for higher microwave ranges and antenna operating in millimeter range (57.5–75.0) GHz. Besides, proposed antenna structures are suitable for applications in beam scanning antennas as well as for forming the desired radiation pattern (for example, cosec^2). Experimentally obtained results are in good agreement with those obtained by simulation.

Acknowledgment

References

1. A. Nesic, I. Radnovic, "High Gain Millimeter Wave Antenna with Cylindrical-Parabolic Reflector", TELSIKS 2009, Nis, Conference Proc. pp. 376-379
2. A. Nesic, I. Radnovic, "New Type of Millimeter Wave Antenna with High Gain and High Side Lobe Suppression", Optoelectronics and Advanced Materials – Rapid Communications Vol. 3, No.10, October 2009, pp. 1060-1064.
3. A. Nesic, I. Radnovic, "60 GHz Range High Gain Printed Antenna Array with a Cylindrical-Parabolic Reflector", Frequenz, 3-4/2010, Vol. 64 March/April 2010, 48-51.
4. A. Nesic and D. Nesic, "Cylindrical-Parabolic Antenna Fed by Printed Axial Array Like Primary Radiator", Patent Pending no. P-207/0738, 24. Sep. 2007
5. A. Nesic, Z. Micic, S. Jovanovic, I. Radnovic , D. Nesic, "Millimeter Wave Printed Antenna Arrays for Covering Various Sector Width", IEEE Antennas and Propagation Magazine, Vol.49, No.1, Feb. 2007, pp. 113-118
6. M. Mikavica, A. Nesic, CAD for Linear and Planar Antenna Array of Various Radiating Elements, ARTECH House, Norwood, MA, 1992
7. J. H. C. van Heuven, "A New Integrated Waveguide-Microstrip Transition", IEEE Transactions on MTT, vol. 24, 1976, pp. 144-147
8. *WIPL-D software package*, http://www.wipl-d.com/, WIPL-D d.o.o.: Belgrade, Serbia, 2012
9. M. Petersson, *Microstrip Solution for Innovative Microwave Feed Systems,* Department of Science and Technology, Linkoping University, Sweden, 2001
10. Rogers Corporation, MWI-2010, Calculator
11. D. Nesic, I. Jokic, M. Frantlovic and M. Sarajlic, "Wide Band-stop Microwave Microstrip Filter on High-resistivity Silicon", MIDEM - Journal of Microelectronics, Electronic Components and Materials, vol. 42, no. 4, 2012, pp. 282-286

Triton X-100 as an Effective Surfactant for Micropump Bubble Tolerance Enhancement

Borut Pečar[1], Drago Resnik[1,2], Matej Možek[1,2], Uroš Aljančič[1,2], Tine Dolžan[1],
Slavko Amon[1,2] and Danilo Vrtačnik[1,2]

[1]University of Ljubljana, Faculty of Electrical Engineering, Laboratory of Microsensor
Structures and Electronics, Ljubljana, Slovenia
[2]Centre of Excellence Namaste, Ljubljana, Slovenia

Abstract: Improvement of bubble tolerance and priming in micropumps based on Triton X-100 surfactant is investigated. Transparent membrane piezoelectric micropumps were fabricated. Precise air volumes were introduced into the micropump chamber to emulate micropump bubble disturbance. Micropump recovery time decreased with increased addition of Triton X-100 surfactant between 50-100 ppm. Effective recovery is mainly a consequence of the air bubble dispersion into a foam of small bubbles. Small bubbles are then readily removed by liquid flow, leading to significant enhancement of micropump bubble tolerance.

Key words: piezoelectric micropump, bubble tolerance, bubble decay, priming, surfactant, Triton X-100

Triton X-100 za izboljšanje zanesljivosti mikročrpalke pri črpanju dvofaznega medija

Povzetek: V prispevku predstavljamo metodo za preprečevanje odpovedi mikročrpalke v primeru črpanja mešanice kapljevine in plina in izboljšanje omočljivosti pri začetnem polnjenju z medijem. Boljša omočljivost prepreči nastanek zračnih žepov pod membrano in omogoči začetni zagon mikročrpalke. Metoda temelji na dodajanju majhnih količin sredstva za zmanjševanje površinske napetosti (Triton X-100) neposredno v medij. Izdelali smo PZT difuzorsko mikročrpalko s stekleno membrano, ki nam je omogočila opazovanje mešanja kapljevine in plina med njenim delovanjem. Dodajanje majhnih količin sredstva za zmanjševanje površinske napetosti v črpan medij (50 ppm ut. Triton X-100) povzorči pod vplivom nihajoče membrane razpad plinskih mehurčkov v množico manjših mehurčkov, ki lažje zapustijo mikročrpalko. Z večanjem koncentracije Tritona (do 100 ppm ut. Triton X-100) smo dosegli hitrejše izločanje plina iz mikročrpalke in močno povečali zanesljivost črpanja.

Ključne besede: piezoelektrična mikročrpalka, dvofazni tok, sredstvo za zmanjševanje površinske napetosti, Triton X-100

* Corresponding Author's e-mail: borut.pecar@fe.uni-lj.si

1 Introduction

Mircoscale pumping technology has attracted a great attention in recent years. With the increasing demand of industrial and medical fields, micropumps have been applied to numerous applications. Some of the potential applications are drug delivery systems, biochemistry, lab-on-a-chip, controlled fuel delivery in engines and fuel cells, localized cooling in electronics, micromixing [1]. Although peristaltic, reciprocating and rotary pumps all ever show up in literatures on micro mechanical driving systems, the reciprocating type micropumps are always in the majority. The most popular reciprocating type micropumps applied in MEMS are piezoelectric, electrostatic, thermo-pneumatic, bimetallic, shape memory alloy (SMA) and ionic conductive polymer film

(ICPF) [2]. The application of a micropump requires maximum reliability, which means that the device should be tolerant towards gas bubbles and able to prime itself. Pumping over long time periods, during which small gas bubbles can be introduced in the pumping liquid, should not degrade the device performance [3]. To prevent entrapment of gas bubbles inside micropump chamber before micropump startup, complicated and unreliable manual priming procedures had to be performed in the first time [3]. A more practical approach was made with a CO_2-purge of the dry device [4]. Residual CO_2 inside the pump was easily dissolved in the following aqueous priming solution, which resulted in a complete filling. The problem of bubbles travelling towards the micropump in the inlet tubing, however, remained unsolved. These problems were discussed quite

early sometimes even with the pessimistic argument that a microdiaphragm pump would not be able to be self-priming at all due to several physical reasons [3]. This discussion was set in 1996 with the first "self-filling" polymer-fabricated micropump by Döpper et al. [4]. A first comprehensive treatment of the subject was performed in 1998 by Richter et al [5]. They proposed increasing of compression ratio in order to improve self-priming and bubble tolerance in reciprocating type micropumps.

An alternative approach for enhancing bubble tolerance and priming in reciprocating type micropumps is proposed. It will be shown that even small quantities of appropriate surfactant additive in liquid medium facilitate priming and significantly enhance bubble tolerance. Presented approach is useful in many specific applications e.g. in living cells manipulation where micropump high-compression ratio and check valves could lead to permanent cells damage. In such micropumps, bubble tolerance can be increased by reducing chamber depth in order to decrease the dead volume and consequently increase the compression ratio. [6]. However, by reducing the chamber depth under a critical point, the deflected diaphragm could damage living cells by compressing them to the chamber floor. Typical diameter of living cells is 10 – 50 μm [7] which dictates the minimum chamber depth. In case of larger cells, obviously deeper chamber is required, deteriorating the compression ratio. In this very specific area, the proposed approach could fill the gap: by adding small amounts of surfactant in cell transportation medium.

In this study fluid was DI water and as surfactant, Triton X-100 (Merck™ / Darmstadt / Germany, $C_{34}H_{62}O_{11}$, 646.9 g/mole, purity 98–100%) was applied. The micropump applications that allow use of Triton as surfactant includes pumping suspensions with microcapsules and localized cooling in electronics. Furthermore, Triton can be added into medium (water) used in initial micropump prothotypes characterisations in order to simplify priming procedure and enhance validity and repeatability of micropump performance measurements.

There are also other appropriate non-toxic surfactants on the market, including e.g. pluronic polyols [8], Tween 80 [9], Pluronic® nF68 [10] and natural surfactants derived from animal sources containing surfactant proteins B (SP-B) and C (SP-C) [11].

2 Experimental

2.1 Micropump fabrication

To get a detailed insight into micropump operation and related bubbles problems, a low-compression ra-

tio diffuser piezoelectric micropump with transparent membrane was fabricated. Micropump consists of 380 μm thick (100) silicon substrate (Fig. 1, A), thinned glass membrane from Pyrex 7740 (Fig. 1, B), piezoelectric actuator (Fig. 1, C) bonded on Pyrex and inlet/outlet stainless steel connections (Fig. 1, D). Micropump fabrication steps are shown on Fig. 2.
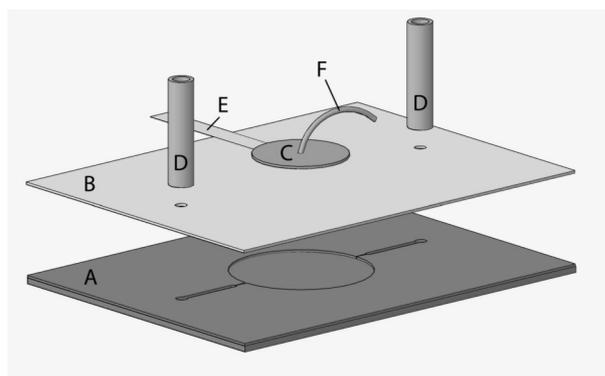
Figure 1: Micropump schematics.
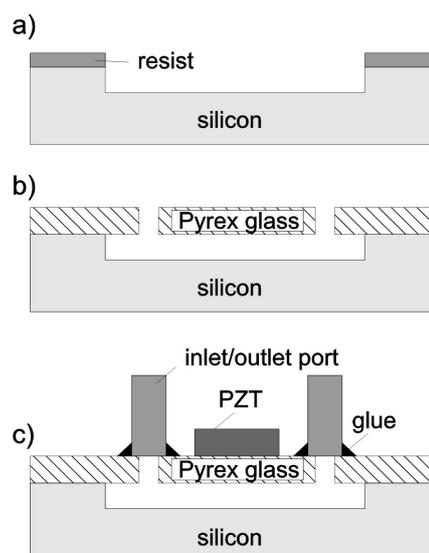
Figure 2: Schematic presentation of fabrication steps: a) mask pattering and DRIE etch of chamber/diffusers/microchannels b) Pyrex cover to Si c) gluing of fluid ports and PZT actuator.

Starting substrate was single-side mechanically polished 380 μm thick (100) silicon wafer. Silicon oxide of thickness 2.5 μm obtained by LPCVD deposition and photoresist AZ 9260, Clariant, USA, of thickness 6.5 μm were used for the fabrication of mask for silicon etching [12]. DRIE system Plasmalab System 100 – ICP 180 based on Bosch process was used for etching deep silicon micropump structures such as chamber, diffuser elements and microchannels (Fig. 2, a). Etched microstructures on silicon chip were finally sealed by anodic bonding of Pyrex glass [13], (Fig. 2, b).

In a previous work [14], it was found that PZT/Pyrex thickness ratio around 1:1 results in good micropump performance (backpressure of 61 mBar and flow-rate of 0.29 ml/min at the resonant frequency of 200Hz). A 200 μm thick PSM23 Noliac piezoelectric disc attached to the glass membrane was used for micropump actuation.

2.2 Characterization methods

To investigate the influence of surfactant additive on micropump bubble tolerance, an appropriate measurement setup was realized (Fig 3). First, micropump flow-rate vs. time characteristic was measured for pure DI water and then measured again for DI water with surfactant addition. To emulate real-world operation conditions, in both cases air bubbles were introduced into the system by a syringe pump.

Figure 3: Measurement setup for bubble tolerance investigation.

For studying liquid/gas interaction in micropump chamber as well as micropump operation after bubble entered the pump, a microcamera for continuous monitoring and a real-time flow-rate measurement method were applied.

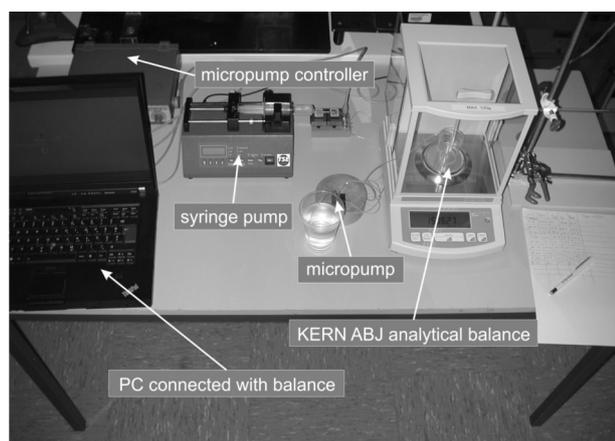In initial micropump characterizations, classic volumetric method was used for micropump flow-rate measurements. However, this method is complex and depends on operator related factors [15], so real-time measurement with high data rate acquisition needed in our case can not be performed accurately.

Therefore, flow-rate measurement by micro weighing method was introduced. Micro weighing method is based on the measurement of the liquid mass change somewhere in the liquid path, measured by a sensitive digital microbalance over the unit of time. Based on our experiences, this approach may be considered as an accurate method for flow-rate measurement since

both mass and the time can be evaluated fast with high precision. For this purpose, KERN ABJ 120-4M analytical balance with RS232 interface was utilized. The balance offered precise measurements up to 120g with 0.1 mg readout. Furthermore, acquired data were stored in a computer with 240 ms sample rate. To minimize dynamic measuring errors caused by liquid dropping into the tank, the output tube was submerged into the tank medium. In order to completely prime the pump, inlet tube was first submerged into the pumping medium while outlet tube was connected to a syringe pump running in the reverse mode.

To introduce controlled quantity of bubbles for bubble tolerance investigation, micropump inlet tube was taken out of the filling tank, by further running syringe pump connected on micropump outlet tube in reverse mode with very low flow-rate setting (cca. 0.05 ml/min). Air bubble volume entering into the liquid system was determined by the filling rate and time. After the air bubble was created in the inlet tube, the syringe pump was disconnected from micropump outlet tube, inlet tube submerged back into the pumping medium and the running micropump itself sucked the air bubble into the chamber. Before micropump flow-rate measuring with microweighing method, air bubble volume can also be determined by measuring air bubble length in micropump inlet tube with known diameter.

Micropump bubble tolerance was investigated for different volumes of introduced air bubbles (1-10 μl) and various concentrations of surfactant additives, expressed as the weight mixture of Triton X-100 / DI water and given in ppm units.

For flow-rate Φ determination, every 240 ms weight measuring data were acquired by a PC what enabled fast flow-rate measuring. Flow-rate Φ is determined from the first derivative of mass m with respect to time:

$$\Phi(t) = \frac{1}{\rho} \frac{dm(t)}{dt} \tag{1}$$

The transparent glass pump cover made it possible to inspect visually micropump priming and surfactant effects in the chamber.

3. Results and discussion

In the following investigations, for micropump actuation sine-wave 200V / 200 Hz excitation signal was applied. As expected, initial measurements with DI water as liquid medium without surfactant additives revealed that bubble tolerance was rather low. Therefore, the influence of surfactant additive on micropump operation such as priming and bubble tolerance was investigated.

3.1 Micropump priming enhancement

Micropump in case of pure DI water was not self-priming, due to the mentioned diffuser valves and low-compression ratio design. It was observed that it is difficult to fill the entire micropump chamber with pure DI water, without entrapping air bubbles. To prime completely the pump, the support of external priming system was required. In most cases, incomplete priming decreased low-compression ratio diffuser micropump performance or even fully prevented micropump start-up.

By addition of Triton X-100 surfactant in DI water, micropump was easily primed. Fluid filled the entire micropump chamber, without entrapping air bubbles. It was found that even small quantities of surfactant additive significantly facilitate micropump priming procedure. Surfactant addition can be used only for initial micropump priming, and can then be removed from the medium.

Therefore, the proposed approach with surfactant addition to the medium results in a simple, reliable and complete priming of the micropump.

3.2 Micropump bubble tolerance enhancement

Micropump bubble tolerance is a critical effect limiting micropump use in many applications. Real-world micropump applications require maximum reliability which is closely related to the micropump insensitivity on gas bubbles in pumping medium. Therefore, effects during bubble traveling through the pump were studied in detail. For this purpose, a transparent membrane micropump was conceived, as described in chapter 2.
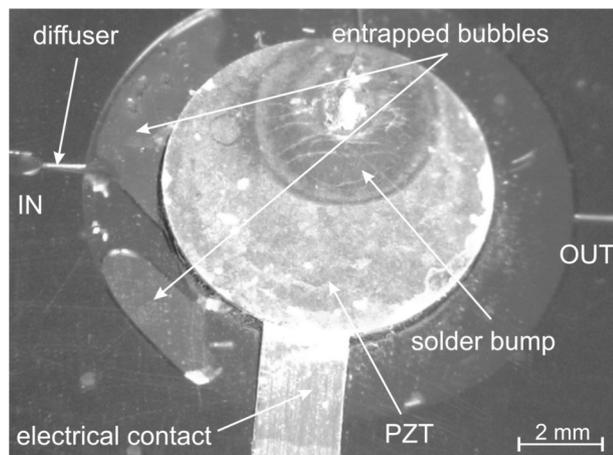

Figure 4: Micropump chamber during micropump actuation (medium: DI-water). Entrapped bubbles (on the left side) prevented further operation.

When a bubble entered into micropump chamber, filled with pure DI water, it split in several smaller bubbles which adhere on the chamber wall. They remained fixed there and due to air compressibility the micropump can not pump out the bubbles by itself (see Figure 4 at the left side).

PZT actuating energy is lost in compressing and decompressing bubbles, what results in significant liquid flow decrease, as can be seen from flow-rate vs. time measurements (Fig. 5, thin line).
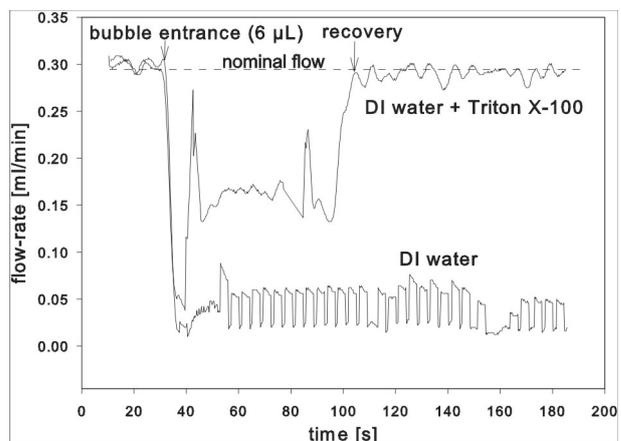
Figure 5: Pump flow-rate vs. time for pure DI water and for mixture of 1000 ppm Triton/DI water. In each case, 6 µL air bubble was introduced.

According to the literature [16], 1000 ppm Triton/DI water mixture has the lowest surface tension coefficient (0.030 N/m). For this reason, this liquid/surfactant mixture was used in our initial investigations to assure low pumping medium surface tension and consequently maximum bubble tolerance enhancement. Later on, also lower Triton/DI water concentrations with higher surface tension coefficients were taken into consideration.

When we introduced a mixture of certain concentration of surfactant in DI water, in this specific case 1000 ppm of Triton, the entrapped air bubble was instantly dispersed into a fine foam of very small air bubbles (see Figure 6, at the left side).

The foam cloud spread from the micropump inlet diffuser into the micropump chamber and consequently, due to air compressibility, micropump flow-rate in the first moment dropped almost to zero (see Fig. 5, bold line, at the bubble entrance point). Then, rapid mixing of the foam with liquid was observed, and fractions of the foam were gradually torn away from the main cloud and transported along the micropump chamber toward micropump outlet. Finally, all the foam was drained out from the system and consequently micropump recovered (see Fig. 5, bold line). Pump recovery

Figure 6: Micropump chamber during micropump actuation (medium: 1000 ppm Triton/DI water mixture). Bubble was dispersed into fine foam (at the left side).

time was defined as the time needed for the pump flow-rate to recover to the nominal flow value within ±5%.

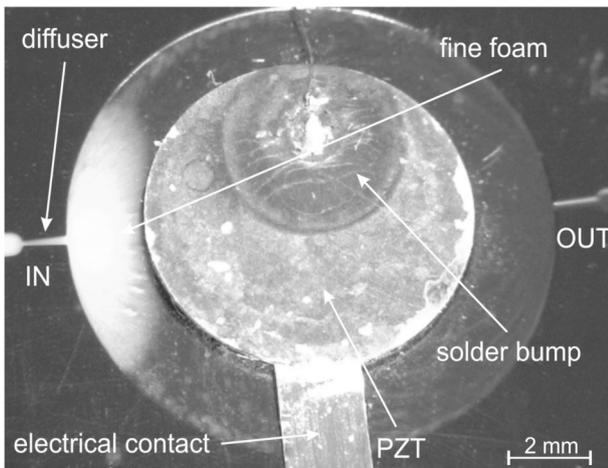Micropump also recovered when subjected to a series of bubbles introduced in the system pumping a 1000 ppm Triton/DI water mixture. The time dependent behavior and recovery of flow-rate due to these sequential disturbances is presented in Fig. 7.
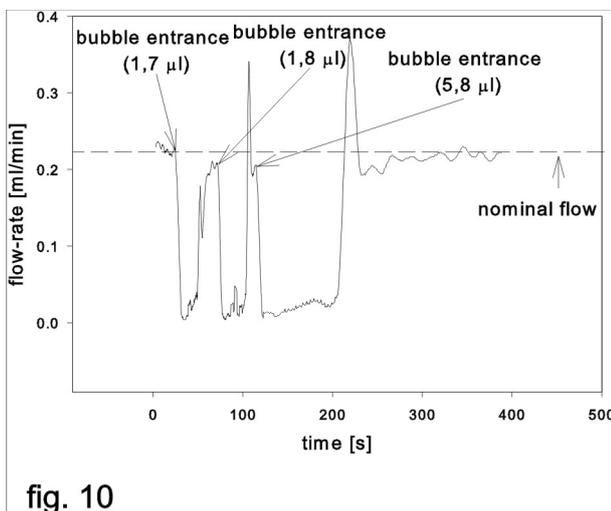

Figure 7: Micropump recovery after a series of 3 sequential bubbles (medium: 1000 ppm Triton/DI water mixture).

Recovery effects were observed also by Richter et al. [1] with bubbles injection in piezoelectrically actuated bubble resistant micropump comprising of improved thinner valve design. Injection of 8 μl air bubble also led to a rapid decrease of pump flow-rate, followed by slow recovery to the baseline within about one minute. The authors described that rather long time interval

needed to recover to the original pump flow-rate had not been understood. Micropump also in this case recovered after a series of bubbles, but when two bubbles were transported into the pump chamber shortly after each other, pump failed.

To investigate micropump flow rate recovery dependence on the volume of air bubble, we introduced in separate experiments three different volumes of air bubbles in the system pumping 1000 ppm Triton/DI water mixture. Recovery time depends in this case strongly on the introduced air bubble volume, as seen in Fig. 8.
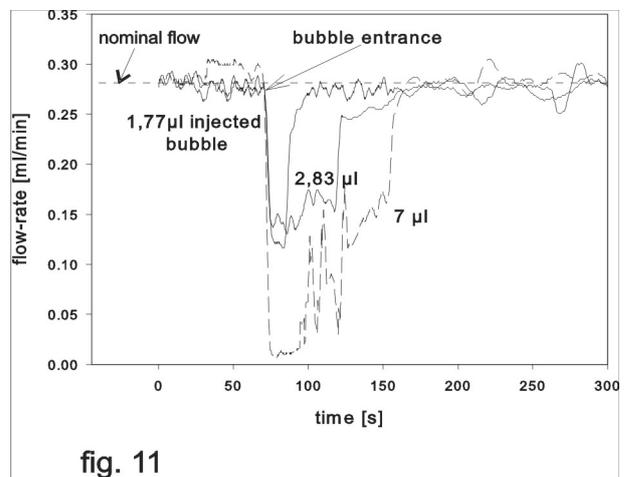
Figure 8: Micropump flow-rate recovery for three different bubble volumes, introduced separately (medium: 1000 ppm Triton/DI water mixture).

To determine micropump recovery time dependency on introduced air bubble volume in a more detail, additional measurements were performed. Due to the transparent micropump diaphragm it was possible to determine the time in which all the foam left the chamber. With a combination of microweighing method for pump flow-rate measuring and optical inspection, it was concluded that immediately after foam was no longer observed in the chamber, micropump flow-rate reached the nominal flow-rate value within ±5%. The time in which the foam had left the chamber therefore determines the recovery time.

Micropump recovery time vs. introduced air bubble volume is shown in Fig. 9. It is seen that two different regimes of recovery time dependencies exists. In both regimes, the recovery dependence was almost a linear function of time, but with lower slope for smaller bubbles (< 7μl) and with higher slope for larger bubbles (> 7 μl).

This implies that two different mechanisms for recovery could be involved. When small air bubbles with

volume up to 7 µl entered the chamber, they dispersed immediately into a fine foam cloud, spreading from micropump inlet diffuser and expanding deep into the micropump chamber. The phenomenon was fully visible through transparent glass diaphragm. Initial foam cloud size in the chamber was proportional to the introduced air bubble size. During micropump operation, foam fractions were gradually removed by the liquid away from the pulsating foam cloud and traveling with the liquid along the micropump chamber toward micropump outlet and finally out of the pump. The amount and velocity of traveling foam fractions were independent of air bubble size. Constant amount and velocity of foam fractions indicated constant foam drain-rate. Therefore, the time in which all the foam had left the chamber (i.e. recovery time) was linearly proportional to the air bubble volume, as seen in Fig. 9 (the part with the lower slope).
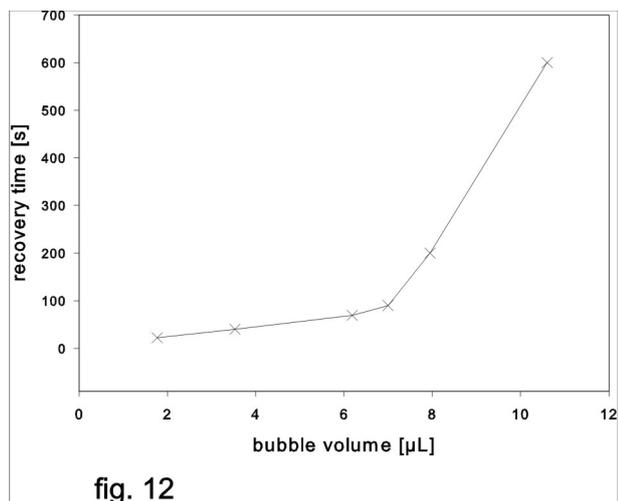

fig. 12

Figure 9: Micropump recovery time vs. introduced bubble volume (medium: 1000 ppm Triton/DI water mixture).

When large bubbles with volumes over 7 µl were introduced in the micropump inlet, an interesting phenomenon was optically observed. Increase of introduced air bubble volume did not additionally increase foam cloud size in the chamber. It seemed that maximum foam amount in the chamber was limited to a dispersion of 7 µl air bubbles and the rest of the bubble remained undispersed at the micropump inlet diffuser, which significantly deteriorated pump performance. At the same time, a reduction in foam fractions removed away from the pulsating foam cloud was observed and fluid velocity declined. Undispersed air bubble was waiting at the inlet diffuser until some of the foam had slowly left the pump chamber. Then a part of undispersed air bubble was transformed into foam etc. When entire air bubble had dispersed into the foam, foam drain-rate increased. And the foam cloud size started to reduce.

When it disappeared, the micropump finally recovered. We assume that the maximum foam amount allowed in the chamber at a time (margin bubble volume of 7 µl, see Fig. 9) is closely linked with the micropump geometry (especially with chamber volume) and Triton/DI water concentration.

3.3 Surface tension influence on micropump recovery

In the last part of this investigation, medium surface tension influence on micropump recovery was studied. For this purpose, various low Triton/DI water mixtures were prepared (100 ppm, 75 ppm, 50 ppm, 25 ppm, 10 ppm). In order to assure complete dissolution of Triton, every dilution was followed by two hours of mixing with magnetic stirrer. Surface tension coefficients of mixtures were not measured, but rather calculated using proposed mathematical model. Liquid surface tension coefficient vs. Triton/DI water concentration model was obtained by curve fitting to four known Triton/DI water concentration data [16], as suggested in [17] (Fig. 10).
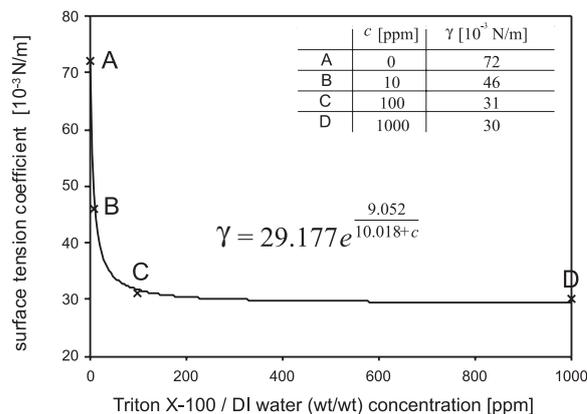
Figure 10: Liquid surface tension coefficient vs. Triton/DI water concentration model [16].

Surface tension coefficients for prepared Triton/DI water mixtures were calculated by using surface tension coefficient vs Triton X-100 / DI water concentration model (Fig. 10):

$$\gamma = 29.177 e^{\frac{9.052}{10.018+c}} \qquad (2)$$

Calculated medium surface tension coefficient exponentially decreased from 0.046 N/m (10 ppm Triton/DI water mixture) to 0.031 N/m (100 ppm Triton/DI water mixture), as shown in Table 1.

Table 1: Calculated surface tension coefficient for various Triton/DI water mixtures.

Triton X-100 / DI water (wt/wt) concentration c [ppm]	Calculated surface tension coefficient γ [10-3 N/m]
10	0.0460
25	0.0378
50	0.0338
75	0.0323
100	0.0310

By introducing surface tension coefficient instead of surfactant concentration, the pump recovery behavior can be easily compared also to other types of surfactants.

Measured micropump flow-rate recovery with respect to different surface tension coefficient and introduced air bubble with 10 μl volume is given in Fig. 11.
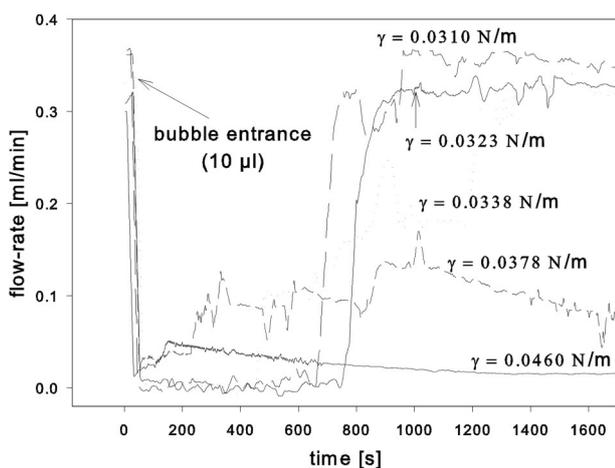
Figure 11: Micropump flow-rate vs. medium surface tension coefficients (10 μl introduced air bubble).

As observed in Fig. 11, micropump recovery begins at a critical medium surface tension of 0.0338 N/m (50 ppm Triton/DI water mixture). Liquid medium surface tension higher than critical surface tension resulted in micropump bubble intolerance. It is assumed that micropump recovery time could be further improved by decreasing medium surface tension below 0.031 N/m which was the limit in our case of Triton/DI water mixtures.

Furthermore, it was observed that also foam structure was dependent on surfactant concentration. At low Triton concentrations, there was rough foam with some larger bubbles which were trapped on chamber wall and permanently deteriorated micropump performance. At higher Triton concentrations, more fine bub-

bles foam texture resulted and consequently higher foam drain-rate and recovery.

Measurements presented here were all taken for micropump zero-load. Additional experiments revealed that surfactant addition also enhanced bubble tolerance in the case of micropump with load pressure on the outlet. For example, it was found that micropump with 33 mBar output load pressure recovered in 82 s when air bubble with volume of 1,9 μl was introduced in the micropump chamber, for 1000 ppm Triton X-100/DI water mixture.

4.   Conclusion

An approach for enhancing micropump priming and bubble tolerance by the application of an appropriate surfactant, such as Triton X-100 is proposed. For detailed investigation of bubble related effects, micropumps with transparent membrane were fabricated by silicon/Pyrex micromachining processing. For emulation of micropump bubble disturbance, precise air volumes were introduced into the micropump chamber. Due to air compressibility the entrapped bubbles in the chamber decreased the micropump performance significantly. By adding a small amount of surfactant in the pumping DI water medium, micropump recovery time was improved. Recovery time was found to be strongly dependent on Triton addition and volume of introduced air bubble. It was determined that application of Triton disperses air bubble into fine foam of small bubbles which do not adhere to the walls and are consequently drained out of the pump by the pumped liquid. After recovery, the micropump performance is completely restored. Regarding enhanced priming, it is shown that only a short initial injection of Triton at pump start-up is sufficient for complete priming.

Acknowledgments

References

1.   Nabavi M 2009 Steady and unsteady flow analysis in microdiffusers and micropumps: a critical review *Microfluid. Nanofluid.* 7 599–619

2.	Tsai N C and Sue C Y 2007 Review of MEMS-based drug delivery and dosing systems *Sensors Actuat. A* 134 555–564

3.	Woias P 2005 Micropumps-past, progress and future prospects *Sensors Actuat. B* 105 28–38

4.	R. Zengerle, M. Leitner, S. Kluge and A. Richter 1995 Carbon dioxide priming of micro liquid systems *proc. Int. Conf. MEMS '95 (Amsterdam, The Netherlands, 29 January–2 February 1995)* pp. 340–343.

4.	Döpper J, Clemens M, Ehrfeld W, Kämper K P and Lehr H 1996 Development of low-cost injection molded micropumps *proc. 5th Int. Conf. New Actuators 1996 (Bremen)* p37

5.	Richter M, Linnemann R and Woias P 1998 Robust design of gas and liquid micropumps *Sensor Actuat. A-phys.* 68 480-6

6.	Andersson H, Wijnggart W, Nilsson P, Enoksson P and Stemme G 2001 A valve-less diffuser for microfluidic analytical systems *Sensors Actuat. B-chem.* 72 259-66.

7.	Campbell A N, Williamson Heyden B R J 2006 Biology: Exploring Life, Boston, Massachusetts: Pearson Prentice Hall, ISBN 0-13-250882-6

8.	McPherson J C Non-Ionic Surfactants in the Treatment of Third Degree Burns 1993 Defense Technical Information Center 35 pages

9.	Arechabala B, Coiffard C, Rivalland P, Coiffard L and RoeckHoltzhauer Y 1999 Comparison of cytotoxicity of various surfactants tested on normal human fibroblast cultures using the neutral red test, MTT assay and LDH release *J. Appl. Toxicol.* 19 163–5

10.	Petri B, Bootz A, Khalansky A, Hekmatara T, Müller R, Uhl R, Kreuter J and Gelperina S 2007 Chemotherapy of brain tumour using doxorubicin bound to surfactant-coated poly(butylcyanoacrylate) nanoparticles: Revisiting the role of surfactants *J. Control. Release* 117 51-8

11.	Ramanathan R 2009 Choosing a Right Surfactant for Respiratory Distress Syndrome Treatment *Neonatology* 95 1-5

12.	Iliescu C 2007 A microfluidic device for impedance spectroscopy analysis of biological samples *Sensors Actuat. B-chem.* 123 168–76

13.	Aljančič U, Resnik D, Vrtačnik D, Možek M and Amon S 2004 Silicon-glass anodic bonding *inform. Midem* 168-73

14.	Pečar B, Penič S, Možek M, Resnik D, Vrtačnik D, Aljančič U and Amon S 2010 Design, modeling, fabrication and characterization of valveless piezoelectric micropumps *proc. Midem 2010 (Ljubljana)* p83

15.	Thomas C, Jones O D and Harry A 1997 Human Factors in Measurement and Calibrations *Measurement Good Practice Guide* vol 8

16.	Reed R L and Taber J J 1964 Gulf Research & Development Co., U.S. 3, 147, 806

17.	Bashir R, Smith E J, Stolle F D 2008 Surfactant - Induced Unsaturated Flow: Instrumented Horizontal Flow Experiment and Hysteretic Modeling *SSSAJ* 72

A dynamic adaptive arbiter for Network-on-Chip

Yanhua Liu[1,2], Jie Jin[2,3], Zongsheng Lai[1]

[1]*Institute of Microelectronics Circuit & System, East China Normal University, Shanghai, China*
[2]*Jiangsu Provincial Key Lab of ASIC Design, Nantong University, Nantong, China*
[3]*Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai, China*

Abstract: Network-on-chip (NoC) is considered as a promising paradigm to overcome the communication bottleneck of future multicore systems. As a basic component in on-chip router, arbiter has a big impact on the performance of router. In this paper, we propose a novel dynamically adaptive arbiter which is based on the round robin mechanism. The proposed arbiter detects buffer status of input ports and changes priorities of the input port dynamically to enhance the performance of the router. Simulation results show that the proposed arbiter can achieve 7.3% improvement in saturation packet injection rate and 13.3% improvement in saturation throughput of NoC on average, when compared with round robin arbiter. Using Synopsys design tools with 0.18-μm technology, implementation results show that a router with the proposed arbiter needs additional 4.8% area compared to a router with round robin arbiter.

Key words: Network-on-chip (NoC), on-chip router, dynamically adaptive arbiter, round robin mechanism,

Dinamično adaptiven razsodnik za omrežje na čipu

Povzetek: Omrežje na čipu (NoC) se smatra za obetajoč zgled pri premagovanju ovir v bodočih večjedrnih sistemih. Razsodnik, kot glavni element na usmerjevalniku na čipu, ima velik vpliv na učinek usmerjevalnika. V članku je predlagan nov dinamično prilagodljiv razsodnik, ki temelji na mehanizmu krožnega dodeljevanja. Predlagan razsodnik zazna status medpomnilnikov vhodov in dinamično spreminja prioritete vhodov tako da poveča učinek usmerjevalnika. Simulacije prikazujejo, da predlagan razsodnik omogoča povprečno 7.3 % povečanje učinkovitosti pri stopnji injekcije nasičenih paketov in 13.3 % povečanje pri prehodu NoC v primerjavi z delovanjem v krožnem dodeljevanju. Pri uporabi načrtovalskega orodja Synopsys v 0.18 μm tehnologiji predlagani razsodnik potrebuje 4.8 % več prostora.

Ključne besede: omrežje na čipu, usmerjevalnik na čipu, dinamično adaptiven razsodnik, mehanizem krožnega dodeljevanja

* Corresponding Author's e-mail: liu8033@163.com

1 Introduction

Traditional bus-based communication architecture has been proved to be a bottleneck that limits the scalability, reusability and reliability of future System-on-Chip (SoC). Networks-on-chip (NoC) has been proposed as a new communication concept that satisfies requirement of the future SoC [1]. NoC provides technologies for generic on-chip interconnection network realized by routers which connects processing elements (PEs) such as ASICs, FPGAs, memories and IP cores [2, 3]. The communication data of PEs are packetized and transferred through the on-chip network. NoC has significant advantages on performance, reusability and scalability compared to traditional bus-based architecture.

On-chip router is the most dominant component of the high-performance NoC. There are three major aspects in design and implementation of a high-performance router: 1) an improving routing algorithm that provides conflict-free paths between input and output ports, 2) an efficient arbiter that authorizes the requiring input ports based on good schedule mechanism, and 3) an optimal buffer distribution that holds more packets in input ports. Although the routing paths are determined by route computation in the router, connection between input and output ports is implemented by arbiter. Arbiter can determine implementation sequence of the routing paths if there exist some conflict requests for the same output port. Buffer can house the incoming packets that cannot be immediately forwarded due

to output port contention or blocking. Increasing buffer slots for heavy-load input ports can improve performance of router. However, more buffer slots will cause more area and power consumption. If heavy-load input ports can be authorized with high priority when they have requests for connecting output port, the pressure on these input port will be decreased and the performance of the router can be improved. So, among the three major aspects, arbiter plays an important role to enhance the router performance.

In this paper, we propose a novel dynamically adaptive arbiter (DAA) which can change the priority of input port dynamically according to its buffer status. Buffer full signal is used as high priority signal of input port. The input port with high priority is allowed to occupy its desired output port prior to other ports, and thus the buffer pressure of the input port can be decreased efficiently. Simulation results show that NoC using proposed arbiter achieves higher performance.

This paper is organized as follows: Section 2 describes related works for arbiters of on-chip router. Section 3 describes details of the proposed arbiter. Section 4 gives experimental and comparison results. Finally, conclusions are made in section 5.

2 Related Works

Many researchers focused on developing various arbitration schemes in order to achieve an efficient allocation and reduce packet latency. A lot of arbiters have been proposed in the routers of computer network such as round robin arbiter [4], fixed priority arbiter [5], lottery arbiter [6], token ring arbiter [7] and so on.

Round robin arbiter treats each input port fairly and guarantees fairness in scheduling. Using round robin arbiter, each input port have an equal chance to own the output port and the starvation problem can be solved. However, round robin arbiter is too fair and may cause low efficiency for some input ports. Fixed priority arbiter always authorizes the requiring input port with the highest priority when requiring contention happens. The input ports with lower priority may rarely be authorized which results in extremely unfair. Lottery arbiter offers input ports certain numbers of lottery as their priority level. Input port with more lotteries has bigger probability to win the output port. However, if the number of lotteries is static, some input ports which have little lotteries may be hardly responded under heavy traffic load. Token ring arbiter cannot guarantee correctness and may miss some requests from different input ports.

All above arbiters can be used in computer network but they are unfit for on-chip router. For on-chip router, resource consumption, average packet latency and complexity of priority strategy should be considered. Recently, some new arbitration methods based on round robin and lottery mechanisms were proposed for the on-chip router. A priority based output arbiter was proposed in [8] which could eliminate the congestion state of NoC. Before arbitrating, the arbiter counted the number of output port requirements for packets in each input port and gave a higher priority to the input port which had more requirements. Zhu et al. [9] presented three new scheduling methods based on round robin mechanism for the on-chip router. The three scheduling methods used different heuristic information to determine the scheduling sequence. However, these arbitration mechanisms needed to detect all packets in input ports and compute their routing paths before delivering them. This resulted in difficult hardware implementation of control logics and totally unsuitable for on-chip router.

Zhang [10] designed a statistic-based lottery arbiter which did not cause starvation problem. However, this arbiter needed many registers, which leaded to a large amount of resource consumption. A customized priority arbiter based on lottery mechanism for the on-chip router was proposed by Wu et al. [11]. The arbitral priorities were customized according to the communication cases among PEs in NoC. This arbiter used static priority which was not fit for dynamic real applications in NoC. Wang [12] improved the lottery arbiter and presented a dynamic lottery arbiter. The dynamic lottery arbiter detected the loads of input ports in every clock cycle and adjusted the priority of each input port dynamically. The proposed arbiter did not work efficiently under the uniform traffic. Even under non-uniform traffics, some heavy-load input ports might always occupy the output ports which caused starvation problem.

3 Design of dynamic adaptive arbiter

In our proposed arbiter, the input buffer full signals are detected as the high priority signals. A high priority will be given to the input port if the buffer of port is full. In order to prevent packet starvation, we design the proposed arbiter based on round robin mechanism. In this section, we first present a typical NoC platform and head-of-line blocking problem. Then, we describe the working principle of the proposed arbiter.

3.1 NoC platform and head-of-line blocking problem

Fig. I shows a typical 2D mesh NoC and the corresponding router architecture. 2D mesh is the most popular

topology for Network-on-Chip which has good scalability. The communication data named packet can be transferred by the on-chip routers and links in 2D mesh NoC. On-chip router is the core component in the NoC and has big impact on system performance [13].

Wormhole router is one of the most commonly used on-chip routers in NoC. It is easy for implementing and suitable for on-chip network. A typical wormhole on-chip router with dedicated buffer per input port is shown in fig. 1 (b). In the wormhole router, each packet is divided into small unit called flit. Head flit proceeds through all the stages while body and tail flits skip route computation and output arbitration stages. Body and tail flits inherit the output port allocated to the head flit. The last flit in a packet, called tail flit, releases the reserved output port that have been reserved by the header flit of that packet, when it departs the current router. The route computation determines deliver direction of the packet according to the destination of the head flit and the routing algorithm. Route computation sends request to the arbiter after determining deliver direction. Arbiter grants the request and connects output port with input port by MUX.
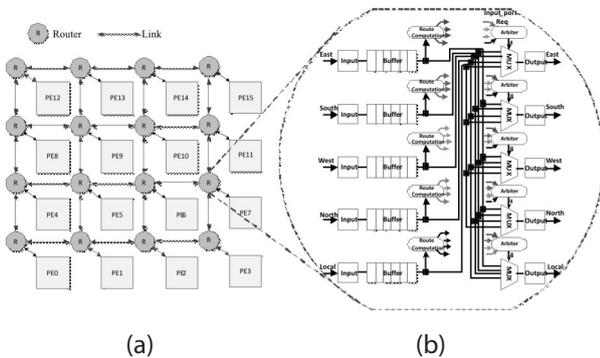

(a)   (b)

Figure 1: (a) 2D mesh NoC topology, (b) wormhole router architecture

Arbiter is one of the key components in the on-chip router. It can determine the output sequence when output contention happens. If packets arrive at different input ports but need to be dispatched into the same output port simultaneously, a output contention will happen. In order to solve the contention, an arbitration mechanism is necessary to allow only one input port to access the output port. Most arbiters are unconcerned with buffer status and authorize the input port by a determinate mechanism. These arbiters may cause head-of-line blocking problems if the input port with full buffer cannot be authorized preferentially. Fig.2 shows an example of head-of-line blocking problem. As shown in fig. 2, west and south input ports request for east output port simultaneously in router R2. If the east output port is connected to the south instead of

west input port in R2, packet p1 in west input buffer of R1 cannot advance. The packets behind p1 will occupy west input buffer in R1 and decrease the network performance. Therefore, head-of-line blocking is a key factor when evaluating different arbiters for on-chip router.
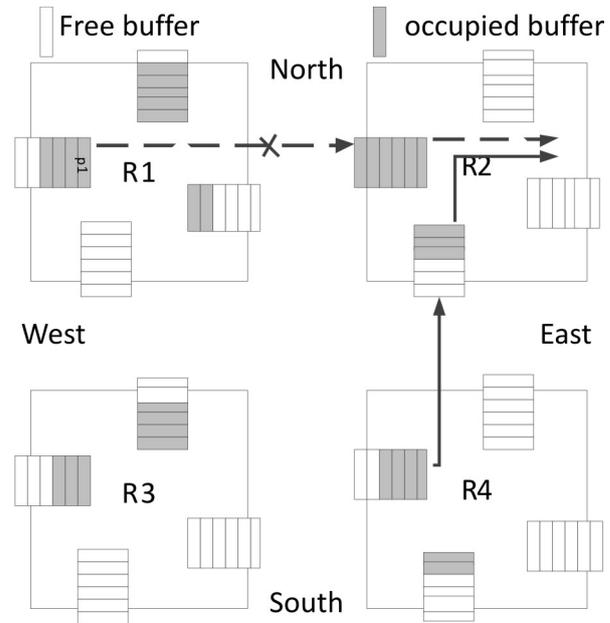

Figure 2: Head-of-line blocking problem induced by inefficient arbitration

3.2 Architecture of dynamically adaptive arbiter

In order to alleviate head-of-line blocking problem, we propose a dynamic arbitration priority for each input port according to the buffer status of the port. If some input ports request the same output port simultaneously, the proposed arbiter will detect the buffer status of these input ports and check whether their buffers are full. Heavy-load input ports makes their buffer easily full. An input port with a full buffer means it cannot hold incoming packets any more. If packets in the full buffer cannot be delivered as soon as possible, the packets will be halted and cause packet latency. For this reason, the input port with full buffer should be authorized with high priority. However, under heavy-load traffic distribution in NoC, there is always more than one input port in a router whose buffer is full. If these input ports request the same output port, these ports should be treated fairly and have an equal chance to win the desired output port. Under the light-load distribution in NoC, there may be no input port in a router whose buffer is full. In other words, no input port has high priority and can be authorized preferentially. So, all input ports should be authorized with an equal chance. Fig. 3 shows a block diagram of one output arbiter in on-chip router which uses the proposed arbitration mecha-

nism. In fig.3, Buffer_full and Input_Port_Req are buffer full signals and request signals from input port of different directions.
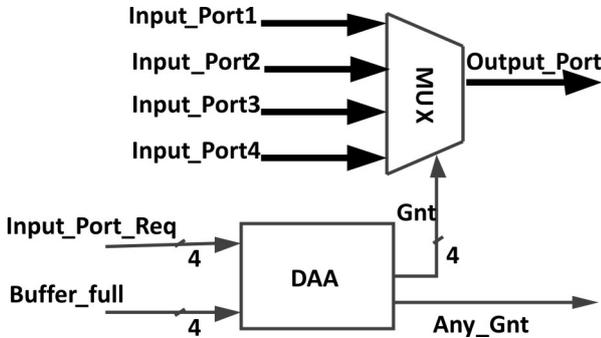

Figure 3: Block diagram for one output port arbiter

Fig. 4(a) gives detailed structure of DAA. As shown in fig. 4(a), DAA is based on two round robin arbiters. In the previous NoC research, many on-chip routers used round robin arbiter as their output arbiter [2] [14]. Round robin arbiter performs well for uniform distribution traffic, but it is not flexible for unbalanced distribution traffics when there are hotspots or priority requirements in NoC. In DAA, if some requests for output ports are generated by the input ports with full buffers, high priorities will be granted to these requests. These input ports will be scheduled first by round robin arbiter1. Other request of input ports with low priorities will be disabled when the high priority ports are scheduled.

In order to prevent starvation for the low priority input ports, a counter and a comparator are used in DAA. The counter is used to record the number of authorizing times for high priority input ports. If the number of authorizing times is bigger than T_threshold, all input ports should be scheduled with equal priority by round robin arbiter2. T_threshold can be customized by NoC designer according to the characteristics of traffics. Moreover, it can be found that, when one of the round robin arbiters works, the request signals for the other round robin arbiter will be disabled. This will ensure that no request signal is missed by DAA in each clock cycle.

Fig. 4(b) shows the typical block diagram of the round robin arbiters. It mainly consists of two barrel shifters, one simple arbiter and one shifter pointer coder. Fig. 4(c) and (d) give the hardware structure of the barrel shifter1 and simple arbiter, respectively. The barrel shifter2 has a similar hardware structure with barrel shifter1 but offers opposite shift direction. The shifter pointer coder generates the shift pointer for two barrel shifters according the previous grant. The detailed working mechanism of the round robin arbiter will not described in this paper because it has been researched
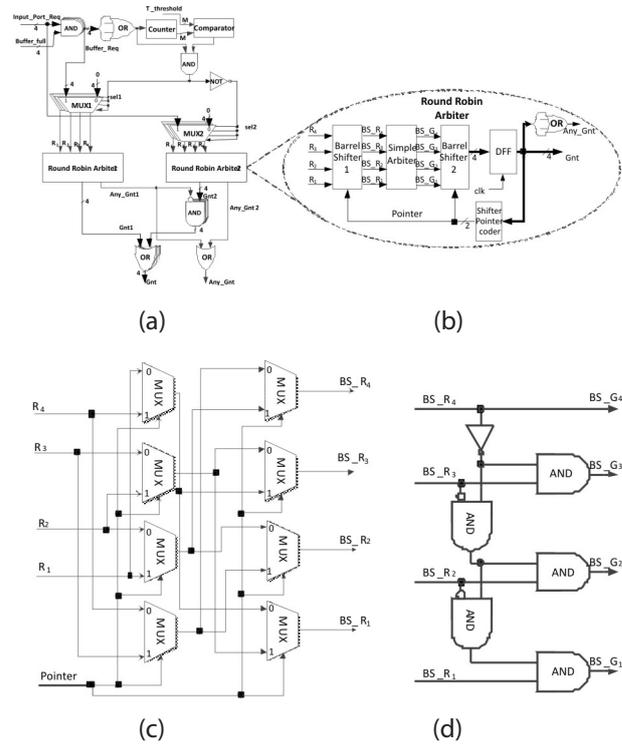

Figure 4: (a) Detailed structure of DAA (b) block diagram of round robin arbiter (c) hardware structure of barrel shifter1 (d) hardware structure of simple arbiter

in many papers [15] [16] [17]. For clarity, the description of input ports scheduling by DAA is illustrated in fig. 5 by pseudo-C language.

Following the scheduling method described in fig. 5, DAA assigns high priorities for input ports whose buffers are full, and priorities are changed dynamically according to the Buffer_full signal. When a buffer of input port is not full, the corresponding buffer full signal will be disabled, and then the input port will lose its high priority. However, with the advance of time, some light-load input ports will accumulate more and more packets if they have no chance to occupy the desired output port. When buffers of these input ports become full, they will be given high priority to connect with their desired output ports. Even for some input ports whose buffers are never full, they will also have a chance to occupy their desired output port by the appropriate T_threshold. This scheduling method avoids the starvation problem thoroughly.

4 Results

4.1 Performance evaluation

In this section, A cycle-accurate NoC simulator implemented in SystemC is used to evaluate the performance

```
input: Input_Port_Req[4], Buffer_full[4], T_threshold[M];

output: Gnt[4], Any_Gnt;

static int count;

Buffer_Req=Input_Port_Req&Buffer_full;

if (Buffer_Req! = 0 & & count < T_threshold)

  { count = count + 1; Gnt = 0;

    for (j = 1; j <= 4; j + +)  / / scheduled by round robin arbiter 1

    {k1=(last_grant_1+j) mod 4; // start from the request which is

    //the first next to last granted one

      if (Buffer_Req[k1]==1) // R_{k1} have a request

      {Gnt[k1]=1; Any_Gnt=1; last_grant_1=k1; break;}

    }

  }

else if ((Buffer_Req==0 || count==T_threshold )&&Input_Port_Req!=0 )

  { if (count==T_threshold) count=0;    Gnt=0;

    for (j = 1;j <= 4; j++) // scheduled by round robin arbiter 2

    { k2=(last_grant_2+j) mod 4;

      if (Input_Port_Req[k2]==1)

      { Gnt[k2]=1; Any_Gnt=1; last_grant_2=k2; break;}

    }

  }

else {Gnt = 0;  Any_Gnt = 0;}  / / no input port is authorized
```

Figure 5: the scheduling method of DAA

of DAA for different traffics. In order to demonstrate the effectiveness of DAA, latency and throughput are chosen as performance metrics of NoC. We compare the performance of NoC based on round robin arbiter [4] (RRA-NoC), lottery arbiter (LA-NoC) [12] and the proposed arbiter (DAA-NoC). All simulations are carried in a 4×4 mesh network for 20000 cycles with X-Y routing and wormhole switching. Four traffic patterns are simulated including three synthetic traffic patterns (Uniform, Bit-complement and Transpose) [18] and one real benchmark (VOPD application) [19]. For synthetic traffic patterns, each packet contains 4~8 flits randomly and for real benchmark, the number of flits in each packet is determined by the bandwidth requirement of the VOPD application.

Latency is defined as the time (in clock cycles) that elapses between the occurrence of a header flit injection into a network at the source node and the occurrence of a tail flit reception at the destination node. We use the average latency, L, as a performance metric like follows:

$$L = \frac{1}{K} \sum_{i=1}^{K} L_i \qquad (1)$$

where K is the total number of packets reaching their destination nodes and L_i is the latency (cycles) of packet i. We define throughput TP as follows:

$$TP = \frac{total\ received\ flits}{number\ of\ PEs \cdot total\ cycles} \qquad (2)$$

where total received flits is the number of flits that arrive at their destination nodes during the simulation, and total cycles is the number of clock cycles elapsed between the start and the end of simulation. Throughput can be used to measure the maximum traffic load that the NoC can handle.

For each traffic scenario, we first give the average packet latency with various packet injection rates (PIRs). Fig. 6 shows the packet latencies for NoCs with different arbiters under the four traffic patterns.

As shown in fig. 6, RRA-NoC, DAA-NoC and LA-NoC perform almost the same when the networks are under light to moderate packet injection rate because there are few full buffers. So, most input ports have an equal chance to occupy the desired output ports. Although some input ports have less chance to win the desired output ports, they have free buffer slots for holding the incoming packets that results in no harm to packet latency. When the networks start to approach saturation, DAA-NoC provides much lower latency compared with RRA-NoC and LA-NoC. For example, when packet injection rate is 0.013 under VOPD pattern, the latencies of RRA-NoC, LA-NoC and DA-NoC are 19 cycles, 18 cycles and 14 cycles respectively. As the packet injection rate is increase to 0.019 under VOPD pattern, the latencies of RRA-NoC, LA-NoC and DA-NoC are 376 cycles, 130 cycles and 91 cycles respectively.

Besides that, The saturation PIR of DAA-NoC also shows an improvement no matter under uniform or unbalanced traffic patterns. For above four traffic patterns, the saturation PIR of DAA-NoC increase by 1.5%, 9%, 7.1% and 11.7% respectively, compared with RRA-NoC. So, the average saturation PIR for DAA-NoC is about 7.3% higher than RRA-NoC. Compared to RRA-NoC, LA-NoC performs worse under uniform traffic pattern, but better under the other three traffic patterns. For uniform traffic pattern, lottery arbiter cannot distinguish the high priority input port clearly. Thus, all input ports win the desired output ports based on luck in LA-NoC. The average saturation PIR of LA-NoC is about 4.2% higher than that of RRA-NoC.

After comparing the latencies and saturation PIRs, fig.7 presents the saturation throughput of different NoCs. Considering the four traffic patterns utilized in this work, the average saturation throughputs of LA-NoC and DAA-NoC increase by 8.5% and 13.3% respectively, compared with RRA-NoC. Even under uniform traffic pattern, the DAA-NoC also achieves about 3.2% saturation throughput improvement compared to RRA-NoC.
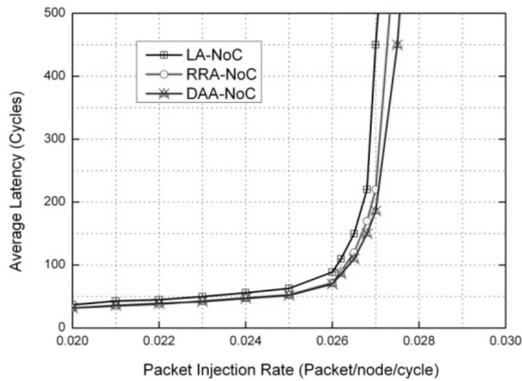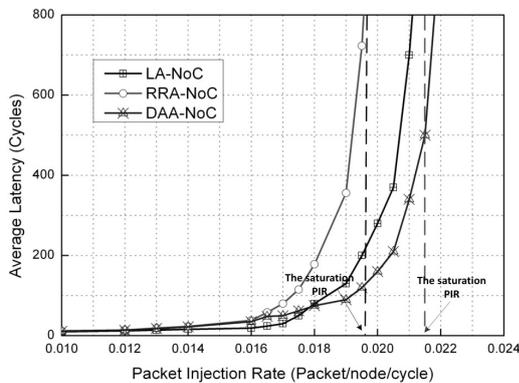
(a)Uniform
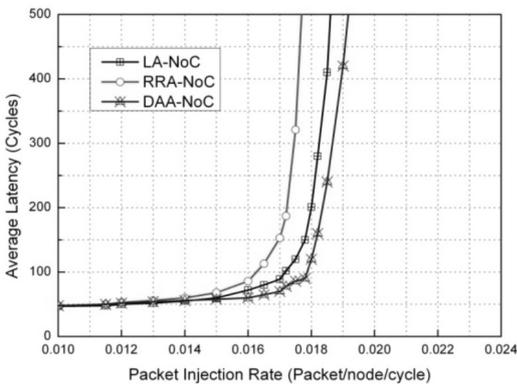
(b) Bit-complement
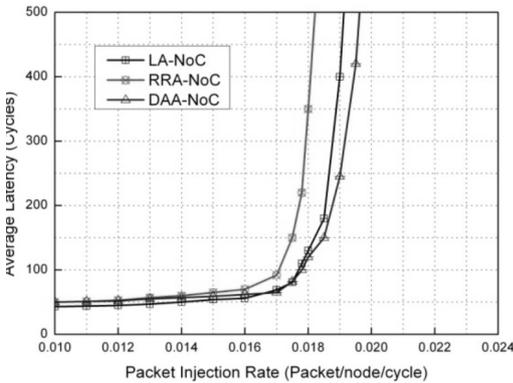
(c) Transpose

(d) VOPD

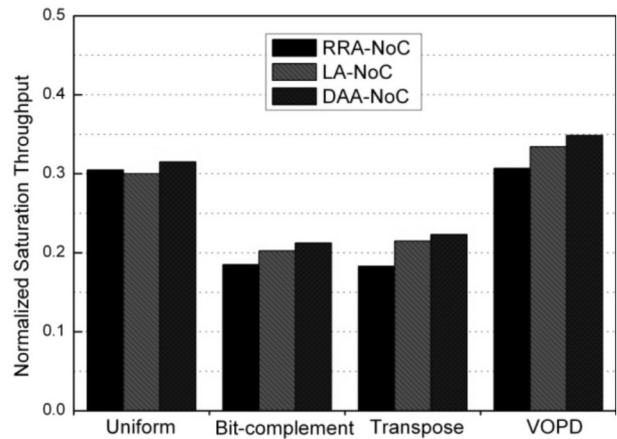**Figure 6:** Average packet latency for NoC with different arbiters

Figure 7: Saturation throughput comparison of NoCs using different arbiters.

4.2 Hardware implementation

Hardware overhead is also an important metric during NoC design. In this section, we first compare the characteristics of round robin arbiter, lottery arbiter and the proposed arbiter. All these arbiters are described in verilog HDL, and synthesized by Synopsys Design Compiler tool with 0.18-μm CMOS library. Table 1 shows the size and the timing results of these designs.

Table 1: Characteristics of different arbiters

	Area (um2)	Size (# of NAND2)	Critical path delay (ns)	Maximum frequency (GHz)
Round robin arbiter	650	72	0.87	1.1
Lottery arbiter	2,735	304	1.38	0.7
Dynamic adaptive arbiter	1,821	202	1.14	0.85

As shown in Table 1, among all the designs, round robin arbiter runs the fastest with its the smallest size. However, as we pointed out and showed before, round robin arbiter is not efficient enough for most communications. Lottery arbiter runs the slowest because of the most complex logic. Compared to round robin arbiter, the size and critical path delay of lottery arbiter increase by 322.2% and 58.6%, respectively. Dynamic adaptive arbiter is better than lottery arbiter, but worse than round robin arbiter. It consumes 180.6% additional hardware resources and causes 31% extra critical path delay compared with round robin arbiter.

Arbiters are small components in a router. Although there are five output arbiters in 2D mesh on-chip router, the area of arbiters makes up only a small portion

of the whole router area. The router using round robin arbiter (RRA-Router), lottery arbiter (LA-Router) and dynamic adaptive arbiter (DAA-Router) are implemented with 200MHz clock frequency. Table 2 gives the detailed area consumptions of these routers.

Table 2: Area overheads for different arbiters in router

design	Area for five arbiters (um2)	Area for the whole router (um2)
RRA-Router	3,250	121,334
LA-Router	13,675	131,762
DAA-Router	9,105	127,192

As table 2 shows, arbiters consume approximately 2.7% to 10.4% of the total area in the different routers. Compared with RRA-Router, LA-Router and DAA-Router require 8.6% and 4.8% additional chip area for the whole router, respectively. However, using DAA-Router instead of RRA-Router, DAA-NoC achieves average 7.3% improvement on saturation PIR and 13.3% improvement on saturation throughput at the cost of additional 4.8% chip area overhead. Thus, the area overhead of the proposed arbiter for on-chip router is acceptable. Compared with LA-Router, DAA-Router pays less area overhead but achieves more performance improvement for NoC. Besides that, the maximum frequencies of these three routers are similar because the critical paths in these routers are determined by the buffer read/write control logics and arbiters have no influence on critical path delay.

5 Conclusions

In this paper, we proposed a dynamic adaptive arbiter based on round robin mechanism. The proposed arbiter detects buffer status of input ports in every clock cycle and adjusts priority of each input port dynamically. It can authorize the input port for transferring data preferentially if the buffer of port is full. Under uniform traffic and non-uniform traffic patterns in NoC, we compared the performance and hardware overhead of NoC based on round robin arbiter, lottery arbiter and the proposed arbiter. The comparison results showed that the proposed arbiter can improve the performance of NoC with an affordable hardware overhead.

Acknowledgments

References

1. L. Benini, G. De Michel, Networks on Chips: A New SoC Paradigm. Computer, Vol. 35, no.1, Jan.2002, pp.70-78.
2. W.J. Dally and B. Towles, Route packets, not wires: on-chip interconnection networks, in Design Automation Conference, 2001. Pro-ceedings , Las Vegas, NV, IEEE Press, 2001, pp. 684-689.
3. P. Guerrier and A. Greiner, A generic architecture for on-chip packet-switched interconnections, in: Design, Automation and Test in Europe Conference and Exhibition 2000. Proceedings, 2000, pp. 250-256.
4. J. Nagle. On packet switches with infinite storage. IEEE Transactions on Communications, vol. 35 no. 4, Apr. 1987, pp.435–438.
5. F. Poletti, D. Bertozzi, L. Benini, A. Bogliolo, Performance Analysis of Arbitration Policies for SoC Communication Architectures, Design Automation for Embedded Systems, Vol. 8, no. 2-3, pp 189-210, 2003.
6. K. Lahiri, A. Raghunathan and G. Lakshminarayana, Lotterybus:A New High-Performance Communication Architecture for System-on-Chip Designs. DAC, Las Vegas, Nevada, USA. 2001. Processings of DAC,200l: 18~22
7. A. Bystrov and A. Yakovlev, Ordered arbiters, Electronics Letters, vol. 35, no. 11, pp. 877-879, 1999.
8. C. Chenghao, T. Kunlin, L. Feipei, T. Shunhung, A Priority based Output Arbiter for NoC Router 2011 IEEE International Symposium on Circuits and Systems (ISCAS), May. 2011, pp. 1928-1931.
9. X.-J. Zhu, H.-B. Zeng, K. Huang, G. Zhang, Round-robin based scheduling algorithms for FIFO IQ switch, IEEE International Conference on Networking, Sensing and Control, 2008, pp. 46-51.
10. Y. Zhang, Architecture and performance comparison of a statistic-based lottery arbiter for shared bus on chip, Proceedings of the Design Automation Conference, Jan. 2005, vol. 2, pp. 1313-1316.
11. C. Wu, H. Li, Y.-B. Li, Z.-M. Yang, Lottery Router: A Customized arbitral priority NoC router. Internal Conference on Computer Science and Software Engineering, Dec. 2008, pp. 411-414.
12. J. Wang, Y.-B. Li, Q.-C. Peng, T.-Q. Tan, A dynamic priority arbiter for Network-on-Chip, IEEE International Symposium on Industrial Embedded Systems, 2009, pp. 253-256.
13. J. Kim et al., A Gracefully Degrading and Energy-Efficient Modular Router Architecture for On-Chip

Networks, in Proc. of 33rd International Symposium on Computer Architecture, 2006, pp. 4-15.

14. D. Bertozzi, L. Benini, Xpipes: a network-on-chip architecture for gigascale systems-on-chip, IEEE Circuits and Systems Magazine, vol. 4 no. 2, 2004, pp. 18-31

15. S.-Q. Zheng, M. Yang, Algorithm-Hardware Codesign of Fast Parallel Round-Robin Arbiters, IEEE Transactions on Parallel and Distributed Systems, vol. 18, no. 1, 2007, pp. 84-95.

16. J.M Jou, Y.-L Lee, An Optimal Round-Robin Arbiter Design for NoC, Journal of Information Science and Engineering vol. 26, 2010, pp. 2047-2058.

17. F. Guderian, E. Fischer, M. Winter, G. Fettweis, Fair rate packet arbitration in Network-on-Chip, 2011 IEEE International SoC Conference (SoCC), 2011, pp. 278-283.

18. W. J. Dally and B. Towles, Principles and Practices of Interconnection Networks . San Mateo, CA: Morgan Kaufmann, 2004.

19. M. Janidarmian, A. Khademzadeh, M. Tavanpour, Onyx: a new heuristic bandwidth-constrained mapping of cores onto tile-based Network on Chip, IEICE Electronics Express, Vol. 6 No. 1, 2009, pp.1-7.

Crystal Controlled CMOS Oscillator for 13.56 MHz RFID Reader

S. M. A. Motakabber, M. I. Ibrahimy

Department of Electrical and Computer Engineering, Faculty of Engineering, International Islamic University Malaysia, Malaysia

Abstract: A design procedure of CMOS integrated crystal oscillator for 13.56 MHz RFID is described in detail by using mathematical and Mentor Graphics VLSI design tools ADK-3. The system is designed by using CMOS 0.18 μm foundry rules and Level-3 transistor model. The frequency stability of the oscillator is created by using piezoelectric crystal. The designed CMOS crystal oscillator can be integrated with the other parts of the RFID reader systems during VLSI design. The computer-generated phase noise is showed -139.5 dBc/Hz at offset of 10 kHz and the power dissipation is 1.25 mW at power supply 2.2V.

Key words: Crystal oscillator, CMOS oscillator, 13.56 MHz RF oscillator, piezoelectric ISO14443

S kristalom krmiljen CMOS oscillator za 13.56 MHz RFID bralnik

Povzetek: V članku je opisan postopek načrtovanja CMOS oscilatorja z integriranim kristalom za 13.56 MHz RFID s pomočjo matematičnega in Mentor Graphics VLSI načrtovalskega orodja ADK-3. Sistem uporablja 0.18 μm CMOS tehnologijo in model tranzistorjev Level-3. Piezoelektričen kristal skrbi za stabilizacijo oscilatorja. CMOS kristalni oscilator se lahko vgradi v ostale dele RFID sistema med načrtovanjem VLSI. Računalniško generiran šum j eprikazan pri -139.5 dBc/Hz pri odmiku 10 kHz in moči 1.25 mW ob napajanju 2.2 V.

Ključne besede: kristalni oscilator, CMOS oscilator, 13.56 MHz RF oscilator, piezoelektričnost ISO14443

* Corresponding Author's e-mail: amotakabber@iium.edu.my

1 Introduction

Radio Frequency Identification (RFID) is used to identify a tagged object by using radio frequency wave. Due to huge potential and robustness nature, the RFID systems have various types of applications such as products chain management systems, access control electronic tickets, fare collection, product labeling, proximity card etc. In fact the heart of the system is a well stable RF source or oscillator. Almost all modern radio communication system is used at least one highly stable radio-frequency source or oscillator for ensuring the reliable communication. A crystal oscillator has the property of generating extremely stable frequency.

An electronic oscillator circuit produces repetitive electric signal from a dc source. The circuit and operation principles of two main types of electronic oscillator (harmonic oscillator and relaxation oscillator) are completely different. The basic structure of a harmonic oscillator is an electronic amplifier of which output is attached with an electronic filter network. The output of the filter network is feedback again into the input of the amplifier. In the beginning when the power supply of the circuit is switched on, the amplifier's output contains only noise. The noise travels through the filtering network is being filtered out. The output (or a portion of the output) is then re-amplified, filtered and feedback repeatedly until it gradually resembles the sinusoidal output. A piezoelectric crystal may take place of the filter network to stabilize the frequency of oscillation, resulting as a crystal oscillator. There are many techniques to implement the harmonic oscillators [1], because there are different ways to design an amplifier and filter network. On the other hand relaxation oscillator produces non-sinusoidal output wave such as a square or saw-tooth waves. This oscillator contains a nonlinear active component like as transistor is used for periodically charging and discharging the energy in a capacitor or inductor. The change of energy in the de-

vice causes abrupt variations on the output waveform and generates non-sinusoidal wave. Like as harmonic oscillator crystal oscillators are often preferred for generating a stable oscillation.

The integrated circuit is more reliable and stable to implement as an amplifier than discrete components amplifier circuit. Therefore, the CMOS circuits are best suitable for design of the active part of the oscillators with a quartz crystal unit. The current mode operations of analog circuits are more suitable for implementing in the CMOS integrated technology. They have a greater gain-bandwidth product than circuits operating in the voltage mode with the same transistors characteristics [2]. Thus current mode operations of analog circuits are suitable for high frequency analysis. The current conveyor is the basic building block for current mode operation. It can be used for realization of negative impedance converters (NICs) with current or voltage controlled negative input resistance. Such NIC circuits have a great gain-bandwidth product and static characteristics whose parameter can be easily modified to the optimal form for oscillator under design.

2 Electrical Model of the Crystal

A piezoelectric (quartz) crystal can be modelled as an equivalent electrical network with low impedance (series) and high impedance (parallel) resonance point spaced closely together as shown in Figure 1.
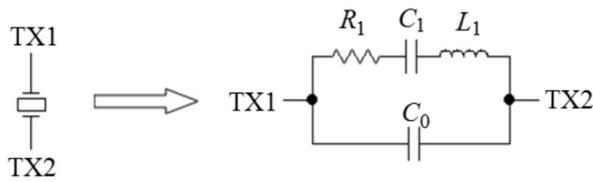

Figure 1: Crystal symbol and its electrical equivalent model

Using Laplace transform, form the equivalent model of the crystal the impedance of this network can be written as:

$$Z(s) = (1/(s.c_1) + s.L_1 + R_1) \| (1/(s.C_0))$$

Or,

$$Z(s) = \frac{s^2 + s.\frac{R_1}{L_1} + \omega_S^2}{s.C_0 \left\{ s^2 + s.\frac{R_1}{L_1} + \omega_p^2 \right\}} \qquad (1)$$

From Equation (1) assume,

$$\omega_s = \frac{1}{\sqrt{L_1.C_1}} \qquad (2)$$

and

$$\omega_p = \sqrt{\frac{C_1 + C_0}{L_1.C_1.C_0}} = \omega_s \sqrt{1 + \frac{C_1}{C_0}} \qquad (3.a)$$

Or,

$$\omega_p \approx \omega_s \left(1 + \frac{C_1}{2C_0} \right); when (C_0 \gg C_1) \qquad (3.b)$$

Where, $s = j\omega$, is the complex frequency, ω_s and ω_p are the series resonant and parallel resonant angular frequency in radians per second respectively. In this research work the design parameters of the crystal $C_0 = 6$ pF, $L_1 = 6.9$ mH, $C_1 = 0.02$ pF and $R_1 = 35$ Ω are considered for generating 13.56 MHz frequency.
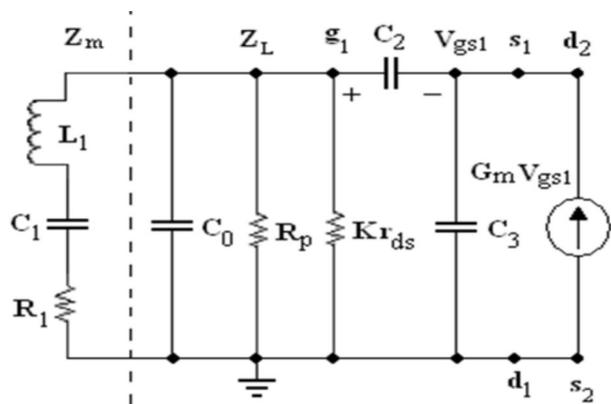
3 Equivalent Circuit of the CMOS Crystal Oscillator

The detailed schematic of the Colpitts crystal oscillator [3] and its equivalent circuit are shown in Figures 2(a) and 2(b) respectively. The nMOS transistor T_1 act as a negative resistance device and transistor T_2 as a bias current source I_b respectively. The transistor T_2 also performed as a current mirror for the reference current I_{ref} in the same chip through nMOS transistor T_5. It provides the stable current with respect to change of the power supply and temperature. The 5pF decoupling capacitor C_4 is added to prevent the high frequency noise leakage from the oscillator. The grid bias resistor R_g combined with the two pMOS transistors T_3 and T_4 is provided the bias for the nMOS transistors T_1 and T_2. The biasing is designated in such a way so that the transistors can always operate in the saturation region during the oscillation. The pMOS transistors T_3 and T_4 are set to be W/L as 0.6/4.0 for giving a bias voltage at the node 1 as $\frac{V_{dd}}{2}$. The external capacitors C_2 and C_3 along with the piezoelectric crystal are worked as a reactive feedback network for three-point Colpitts oscillator circuit.

The capacitor C_2 and C_3 are selected as 10pF and 60pF respectively in this project. The MOSFET's parameters are used in this design as shown in Table 1. In Figure 2(b) the equivalent circuit parameter, $K = \left(1 + \frac{C_3}{C_2} \right)^2$, the effective drain resistance of T_1 and T_2 as, $r_{as} = (r_{ds1} \| r_{ds2})$, and finally $R_p = R_c + (r_{ds3} \| r_{ds4})$ are used. The drain resistance of T_5 is considered as zero value since during the oscillation the capacitor C_4 becomes short circuit.
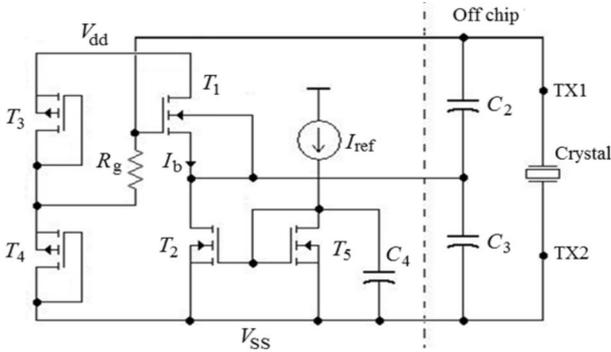
(a)

(b)

Figure 2: (a) Schematic of a Colpitts crystal oscillator and (b) its equivalent circuit

4. Critical Transconductance g_m

The critical transconductance g_m of the transistor T_1 is the minimum value which is essential for sustaining the oscillation of the circuit. Figure 2(b) is the simplified small signal equivalent circuit of the crystal oscillator where the passive motional impedance Z_m is considered as the series $R_1 L_1 C_1$ tank (resonant) circuit. The remaining part of the circuit which includes the passive as well as active components is considered as the load impedance Z_L. On the basis of negative-resistance model of the oscillator, the oscillation may occur [4][5] only if,

$$R_1 + R_e\{Z_L(j\omega_0)\} \leq 0 \tag{4}$$

Here, $R_e\{Z_L(j\omega_0)\}$ is the real component of the impedance Z_L at angular frequency ω_0 of oscillator.

If it is considered that, $R_1 = R_e\{Z_L(j\omega_0)\}$ and $\omega_0 = 2\pi \times 13.56 \times 10^6$ rad/s, the critical transconductance of the transistor T_1 is calculated as $g_m \approx 1.4$ mA/V. Using a safety factor of about 4, the transistor T_1 must have a critical $g_m \approx 5.6$ mA/V. From MOSFET theory, the relation of g_m with transistor's physical dimensions and process parameters are as follows

$$g_m = \sqrt{2\mu_n C_{0x}\left(\frac{W}{L}\right)I_d} \tag{5}$$

The parameters, μ_n and C_{0x} are defined by the process considered, W and L are the physical dimension of width and length of the transistor respectively. The parameter I_d is the bias current of the transistor T_1, in this design it was considered as 150 μA. The bias current I_d can be calculated from the reference current I_{ref}, in this design I_{ref} has been considered as 300 μA. The aspect ratio $\frac{W}{L}$ and the other parameters of the respective transistors are shown in Table 1.

Table 1: Physical dimensions and parameters of the MOSFETs are used in the design. (Technology used 0.18 μm and process used for transistor model Level-3)

FET No.	FET type	W (μm)	L (μm)	Aspect ratio $\frac{W}{L}$	gm (mA/V)	rds (MΩ)
T1	nMOS	13.0	0.2	65.0	5.600	0.01
T2	nMOS	1.0	0.2	5.00	0.450	0.1
T3	pMOS	0.6	4.0	0.15	0.008	50.0
T4	pMOS	0.6	4.0	0.15	0.008	50.0
T5	nMOS	2.0	0.2	10.0	0.850	0.04

5 Estimation of the Oscillation Frequency by the Feedback Model

According to the feedback theory, a circuit would be oscillating only, if the small signal close-loop gain of the circuit is greater than unity and the phase shift of the feedback loop is equal to zero (positive feedback). The closed-loop gain can be represented as Equation (6).

$$T(s) = A(s)F(s) \tag{6}$$

Here, A(s) is the gain without feedback and F(s) is the feedback factor. From Figure 2(b), the value of the generated frequency can be calculated by Equation (7)

$$\omega_0 = \left\{\frac{1}{(L_1 \cdot C_1)} + \left[L_1 \cdot \left(\frac{C_0 + C_2 \cdot C_3}{C_2 + C_3}\right)\right]^{-1}\right\}^{\frac{1}{2}} \tag{7}$$

When the capacitors and inductors design values are used in Equation (7), the generated frequency of oscillation is, $\omega_0 = 85.184 \times 10^6 \frac{rad}{s}$, or, $f_0 = 13.557 \times 10^6 Hz$.

6 Layout Design

The layout design is done in analog design mode by Mentor Graphics design tool kit ADK-3. The designed oscillator layout with an isolation buffer amplifier is shown in Figure 3.
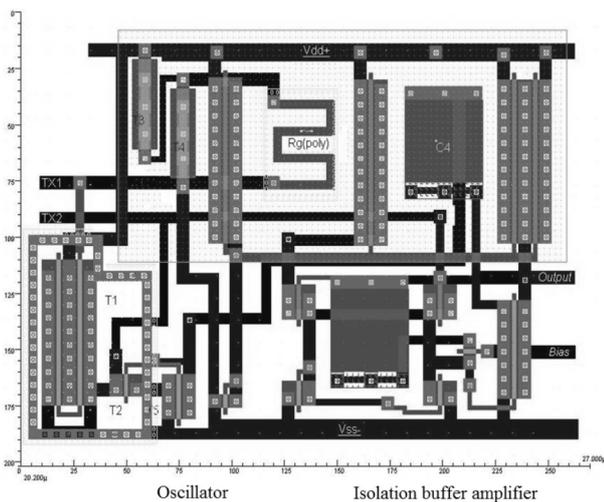

**Figure 3:** The designed oscillator and an isolation buffer amplifier layout

To prevent the oscillator circuit from the substrate noise all the p-channel transistors are placed inside the n-well and the n-well is connected with the power supply rail V_{dd} In addition the n-channel transistors T_1 and T_2 are surrounded by two guard rings which are connected to the substrate. The unused areas of the chip are filled with extra connection to the substrate and well-regions. To avoid the loading effect of the oscillator an isolation buffer amplifier is also designed on the same silicon chip. The oscillator including an analog buffer amplifier is occupied a die area of 27μm x 20μm. The amplitude of the oscillation is controlled carefully to prohibit the potential of the drain of T_1 and the source of T_2 from exceeding power supplies V_{dd} and V_{ss} respectively to prevent the latch-up effect.

7 Simulated Results

The simulated results are shown in Figure 4. The oscillator layout circuit is simulated which has been designed by using CMOS 0.18 μm foundry rules together with a buffer amplifier as shown in Figure 3. Different VLSI design software are used for simulation purpose and it found the same performance of the circuit.

The maximum amplitude swing of the sinusoidal wave is 712 mV$_{PP}$ at power supply 2.2V, and power consumption is 1 25 mW The phase noise measured by simula-

tion is −139.5 dBc/Hz at 10 kHz offset frequency. The designed oscillator can start up reliably within a wide ranges of supply voltage (0.9~3.6V) and able to operate in a wide ranges of temperature (−10~65°C) to maintain a stable frequency.
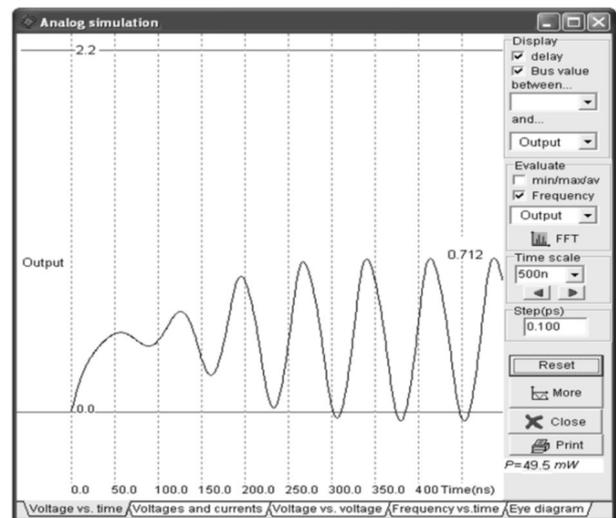
(a)
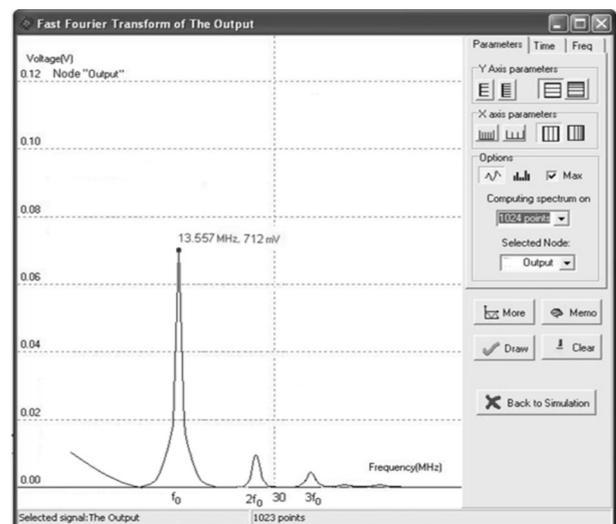
(b)

Figure 4: (a) Build-up of oscillation and time domain signal of the oscillator, (b) frequency domain representation of the output signal

8. Discussion and Conclusion

A step by step procedure of integrated crystal oscillator design has been described. This design can be a guideline for a reliable short time start-up and low phase noise allowing frequency stability ± 7.0 kHz for 13.56 MHz RFID system. In ISO14443 standard this is essential for 13.56 MHz RFID system. Any additional capacitor

across the piezoelectric crystal causes the parallel resonance to shift downward. This technique can be used to adjust the oscillator frequency exactly at 13.56 MHz. The capacitors C_2 and C_3 values are affected the gain of the oscillator circuit and observed that lower the values higher the gain, again the gain is also affected by the ratio $\frac{C_3}{C_2}$, higher the value as the result of higher gain. A buffer circuit is used in this design with the oscillator to derive a load during simulation. This buffer is used to isolate the load from the oscillator circuit which ensures the stable oscillation.

References

1. W. Thommen, "An Improved Low Power Crystal Oscillator", Proceedings of the 25th European Solid-State Circuits Conference, ESSCIRC '99, 21-23 Sept. 1999, pp. 146 -149.
2. I. I. Ivanisevic and D. M. Vasiljevic, "The Quartz Crystal Oscillator Realization Using Current Conveyors", IEEE Transaction on Circuits and Systems-I, Aug. l993, vol. 40, pp. 530-533.
3. P. Andreani, W. Xiaoyan, L. Vandi, and A. Fard, "A Study of Phase Noise in Colpitts and LC-tank CMOS Oscillators", IEEE Journal of Solid-State Circuits, vol. 40 (5), May 2005, pp. 1107-1118.
4. M. Toki and T. Huchisawa, "A Method of Analyzing Negative Resistive and Equivalent Capacitance of CMOS Crystal Oscillators", Electronics and Communications in Japan (Part II: Electronics), vol. 77(6), June 1994, pp. 72-80.
5. B. Meskoob and S. Prasad, "Loop-Gain Measurement and Feedback Oscillator Design," IEEE Microwave and Guided Wave Letters, vol. 2(9), Sept. 1992, pp. 375-377.

Simulation of semiconductor bulk trap influence on the electrical characteristics of the n-channel power VDMOS transistor

Sanja Aleksić, Biljana Pešić and Dragan Pantić

Department of Microelectronics, Faculty of Electronic Engineering, University of Niš, Niš, Serbia,

Abstract: In this paper the impact of traps generated in semiconductor bulk due to High Electric Field Stress (HEFS) or irradiation of n-channel power VDMOSFET are presented. The influence of semiconductor bulk traps is expected, due to the fact that the current mainly flows vertically through the n-epitaxial layer and n+ substrate to drain contact. For the reverse engineering of the process flow and the simulation of electrical characteristics of power VDMOS transistor, Technology Computer-Aided Design (TCAD) software package tools from Silvaco are used. Taking the advantage of simulation, the influences of donor (DT) and acceptor (AT) traps generated in the semiconductor bulk on the electrical characteristics are separately analysed and discussed.

Key words: TCAD, power VDMOSFET, donor and acceptor bulk traps

Simulacija vpliva pasti v substratu n kanalnega močnostnega VDMOS tranzistorja na njegove električne lastnosti

Povzetek: V članku so predstavljeni vplivi pasti v substratu n kanalnega močnostnega VDMOS tranzistorja zaradi sevanja ali vpliva visokega električnega polja (HEFS). Vpliv pasti se pričakuje zaradi vertikalnega toka preko epitaksijske plasti n in n+ substrata v kontakt. Za analizo električnih lastnosti tranzistorja je bila uporabljena TCAD programska oprema proizvajalca Silvaco. Simulacije omogočajo, da je vpliv akceptorskih in donorskih plasti obravnavan ločeno.

Ključne besede: TCAD, močnostni VDMOSFET, donorske in akceptorske pasti v substratu

* Corresponding Author's e-mail: sanja.aleksic@elfak.ni.ac.rs

1 Introduction

The advanced generation of power VDMOS transistors are designed for a wide range of switching and amplifying applications where high breakdown voltage, high input impedance, low input capacitance, and fast switching speeds are desired. These components are widely used in various applications. Considering the facts that the design of highly reliable, high-speed and low power integrated circuits (IC) is the critical task in this procedure, it is very important to know how the power VDMOS transistor acts when it is exposed to various stresses, such as high electrical field stress or irradiation. In this cases the defects in the form of traps (which can be neutral or charged), are generated in ox-ide and semiconductor bulk, as well as at the Si/SiO$_2$ interface. They significantly influence on the electrical characteristics of semiconductor devices. The traps, generated at the Si/SiO$_2$ interface and in the gate oxide, have the dominant influence on the electrical characteristics of MOS transistors [1-7], while in the power VDMOS transistor the influence of generated traps in semiconductor bulk must be taken into account due to the fact that the current mainly flows vertically through the n-epitaxial layer and n$^+$-substrate to drain contact. Further reducing in VDMOS transistor size continually complicates the device physics and makes device modeling more sophisticated [8-10]. Because of that, complete fabrication process flow and device electrical characteristics simulation programs, as well as the

electronic circuit simulators, are the essential tools in the procedure of ICs design.

Over the last thirty years a number of papers and reports dealing with the degradation of the electrical characteristics of semiconductor devices due to different stress conditions have been published, where the physical models for the instability explanations has proposed [11-15]. In the most of these models, the presence of defects in semiconductor substrates was neglected, because in standard MOS structures the defects generated in the gate oxide and the Si/SiO$_2$ interface have the dominant influence on the electrical characteristics. In VDMOS structures their influence could not be ignored, because of vertical current flow. Using the possibilities provided by TCAD simulation tools for separation of different parameters influence, models and mechanisms on the device electrical characteristics, in this paper only the effects of semiconductor bulk traps on the electrical characteristics of n-channel power VDMOS transistor are investigated. The basic data necessary for the power VDMOSFET electrical characteristics simulation is the net doping profile distribution in its two-dimensional (2D) simulation domain. It is determine by the procedure of reverse engineering within the values of basic process parameters are determined on the basis of available information obtained from literature, data sheets and measurements, and the simulation of the complete power VDMOSFET production process flow by using ATHENA process simulator [16]. Afterwards, the electrical characteristics of VDMOS transistors and the influence of bulk trap generation mechanisms on them are simulated by using the device simulator ATLAS [17]. The semiconductor bulk trap mathematical model incorporated in the device simulator ATLAS is given in the second part of this paper. Finally, the obtained simulation results are analyzed, where the impacts of donor-like and acceptor-like semiconductor bulk traps are considered separately.

2 Semiconductor bulk trap mathematical models

The presence of defects in semiconductors (impurities, vacancies, dangling bonds, etc.) has a significant impact on the device electrical characteristics, especially when these defects are charged. These traps are changing the density of space charge and the potential distribution in the device structure and also have the influence on the recombination statistics and carriers mobility. The fact is that the amount of bulk and interface charged traps increases significantly when the devices are exposed to high electric field or radiation [3,

8], and in these cases, accurate simulation of the electrical characteristics of semiconductor devices requires to take into consideration the influence of space charge that comes from stress induced charge traps.

In solid-state physics there are three different mechanisms which add space charge directly into the right hand side of Poisson's equation in addition to the ionized donor and acceptor impurities, and these are interface fixed charge, interface trap and bulk trap states. Interface fixed charge Q$_f$ is modeled as a sheet charge at the Si/SiO$_2$ interface and therefore is controlled by the interface boundary condition, while interface traps Q$_{IT}$ and semiconductor bulk traps Q$_{BT}$ add space charge directly into the Poisson's equation:

$$div(\varepsilon \nabla \varphi) = q(n - p - N_D^+ + N_A^-) - (Q_{IT} + Q_{BT}) \quad (1)$$

This section describes the definition of bulk trap states mathematical models that are implemented in program ATLAS that provides general capabilities for physically-based 2D and 3D simulation of semiconductor devices [17].

2.1 Bulk trap implementation into Poisson equation

Associated energy of bulk traps lies in the forbidden gap and exchange charge with the conduction and valence band through the emission and capture of electrons and holes, as it is shown in Fig. 1. They can be donor-like trap (DT) or acceptor-like trap (AT). DT filled with electron is neutral and with the release of an electron becomes positive charged (ionized). Unlike donors, the energy level for DT lies in energy gap near the top of valence band. Contrary, empty AT is neutral and becomes negatively charged (ionized) when filled with an electron. His energy level lies near the bottom of conduction band [17].

The net charge due to the presence of traps in semiconductor bulk is added on the right hand side of Poisson's equation. The total space charge is defined as:

$$Q_{BT} = q(N_{DT}^+ - N_{AT}^-) = Q_{DT}^+ - Q_{AT}^- \quad (2)$$

N$^+_{DT}$ and N$^-_{AT}$ are the densities of ionized DT and AT, respectively, and their densities are equal to the product of the donor-like trap densities N$_{DT}$ and acceptor-like trap densities N$_{AT}$ in the semiconductor bulk and its probability of ionization F$_{DT}$ and F$_{AT}$.

$$N_{DT}^+ = N_{DT} \cdot F_{DT} \quad (3)$$

$$N_{AT}^{-} = N_{AT} \cdot F_{AT} \qquad (4)$$

The probability of ionization assumes that the capture cross sections are the constant for all trap energy levels in the forbidden band, and follows the analysis developed by Simmons and Taylor [18]:

$$F_{DT} = \frac{v_p \sigma_p p + e_{nD}}{v_n \sigma_n n + v_p \sigma_p p + e_{nD} + e_{pD}} \qquad (5)$$

$$F_{AT} = \frac{v_n \sigma_n n + e_{pA}}{v_n \sigma_n n + v_p \sigma_p p + e_{nA} + e_{pA}} \qquad (6)$$

σ_n and σ_p are electron and hole capture cross sections, respectively, v_n and v_p are the thermal velocities for electron and hole, while the electron and hole emission rates for DT, e_{nD} and are e_{pD} defined by:

$$e_{nD} = \frac{1}{D.FAC} \cdot v_n \cdot \sigma_n \cdot n_i \cdot \exp\left(\frac{E_T - E_i}{kT_L}\right) \qquad (7)$$

$$e_{pD} = D.FAC \cdot v_p \cdot \sigma_p \cdot n_i \cdot \exp\left(\frac{E_i - E_T}{kT_L}\right) \qquad (8)$$

E_i is the intrinsic Fermi level position, E_T is the energy of the discrete trap level, and T_L is the lattice temperature. Parameter D.FAC takes into account the fact that defects in "empty" or "filled" conditions have different spin and degeneracy.

The emission rates for AT, e_{nA} and e_{pA} are defined by:

$$e_{nA} = D.FAC \cdot v_n \cdot \sigma_n \cdot n_i \cdot \exp\left(\frac{E_T - E_i}{kT_L}\right) \qquad (9)$$

$$e_{pA} = \frac{1}{D.FAC} \cdot v_p \cdot \sigma_p \cdot n_i \cdot \exp\left(\frac{E_i - E_T}{kT_L}\right) \qquad (10)$$

In the case when several different donor and/or acceptor trap energy levels are defined, the net space charge is the sum of charges originated from all defined traps.

In order to activate semiconductor bulk trap model in ATLAS device simulation tool and to analyze DT and AT impact on electrical characteristics of semiconductor device TRAP statements is used. The accurate simulation of device physics requires the use of transient trap simulation model, since DT and AT do not reach the equilibrium distribution instantly, but require time for electrons to be emitted or captured. However, this simulation is time consuming and therefore the static model is often used.
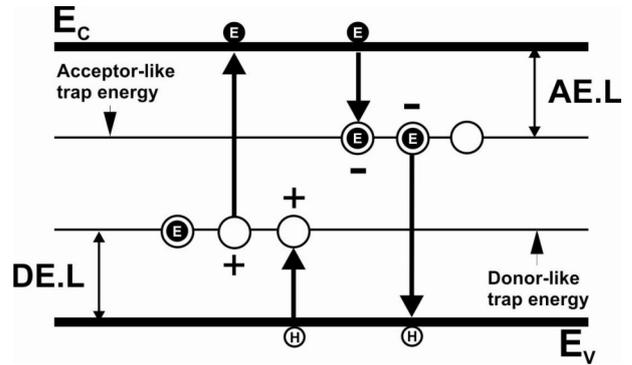

Figure 1: Donor and acceptor trap energy levels and charge states in Si forbidden gap.

2.2 Bulk trap implementation into recombination model

The presence of defects in the semiconductor bulk can significantly affects on the concentration of carriers because that electrons are being emitted or captured by DT and AT. This is accounted in the carrier continuity equations by modifying the standard Shockley-Read-Hall recombination term as follows [19]:

$$R_{SRH} = \sum_{i=1}^{A} R_{Di} + \sum_{i=1}^{B} R_{Aj} \qquad (11)$$

where A is the number of DT energy levels in the forbidden gap, B is the number of AT energy levels in the forbidden gap, while their recombination terms R_D and R_A are:

$$R_D = \frac{pn - n_i^2}{\tau_p\left[n + \frac{1}{D.FAC} \cdot n_i \cdot \exp\left(\frac{E_{TD} - E_i}{kT_L}\right)\right] + \tau_n\left[p + D.FAC \cdot n_i \cdot \exp\left(\frac{E_i - E_{TD}}{kT_L}\right)\right]} \qquad (12)$$

$$R_A = \frac{pn - n_i^2}{\tau_p\left[n + D.FAC \cdot n_i \cdot \exp\left(\frac{E_{TA} - E_i}{kT_L}\right)\right] + \tau_n\left[p + \frac{1}{D.FAC} \cdot n_i \cdot \exp\left(\frac{E_i - E_{TA}}{kT_L}\right)\right]} \qquad (13)$$

E_{TD} and E_{TA} are the donor and acceptor trap energies, E_I is the intrinsic Fermi level and T_L is the lattice temperature. The electron and hole lifetimes τ_n and τ_p are related to the carriežr cross sections σ_n and σ_p through equations:

$$\tau_n = \frac{1}{\sigma_n \cdot v_n \cdot N_{DT}} \qquad (14)$$

$$\tau_p = \frac{1}{\sigma_p \cdot v_p \cdot N_{AT}} \qquad (15)$$

v_n and v_p are the thermal velocities for electron and holes, respectively.

3 Simulation results

In this section the impacts of DT and AT traps in semiconductor bulk on the electrical characteristics of power n-channel VDMOS transistor from IRF510 serial are presented. This device is intentionally used, since its drain current after the channel, flows vertically through the n-epitaxial layer and substrate (Fig. 2.), and therefore the impact of the semiconductor bulk traps on the device electrical characteristics is more pronounced. The simulations have been carried out by using the process simulator ATHENA [16] and the device simulator ATLAS [17], which are the integral part of Silvaco TCAD software package. The reverse engineering of the complete production process flow of the power VDMOS transistor is made by the using of the available technical documentation and its measured electrical characteristics.

3.1 Reverse engineering of power VDMOSFET's process flow

TCAD study of semiconductor bulk traps influence on the electrical characteristics is done on the n-channel power VDMOS transistor IRF510. This transistor is the third advanced generation HEXFETs from International Rectifier, designed for a wide range of switching and amplifying applications where high breakdown voltage, high input impedance, low input capacitance, and fast switching speeds are desired. For the simulation of its electrical characteristics the doping profile distribution in the 2D simulation domain is required. Determination of the doping profile is a serious problem due to the complex structure of the device (Fig. 2) and the fact that the complete technology production flow of power VDMOS transistor has more than hundred process steps, where each process has several parameters that are unknown.

At the beginning of the process simulation the initial data that we know are: the basic information about used technology (substrate concentration epitaxial layer concentration and thickness, source/drain and well regions junction depths and sheet resistances, and the order of process steps in the production flow), design rules for the given technology, information from IRF510 data sheet [20] (threshold voltage V_{TH}, gate oxide thickness channel length, drain to source on-resistance transfer and output electrical characteristics) and measured electrical characteristics of power VDMOS transistor ($V_{TH} = 2.7$ V, $I_D = f(V_{GS})$, $I_D = F(V_{DS})$).

The production process flow is reconstructed by using all this data and information about device geometry, and the values of process parameters such as: implan-
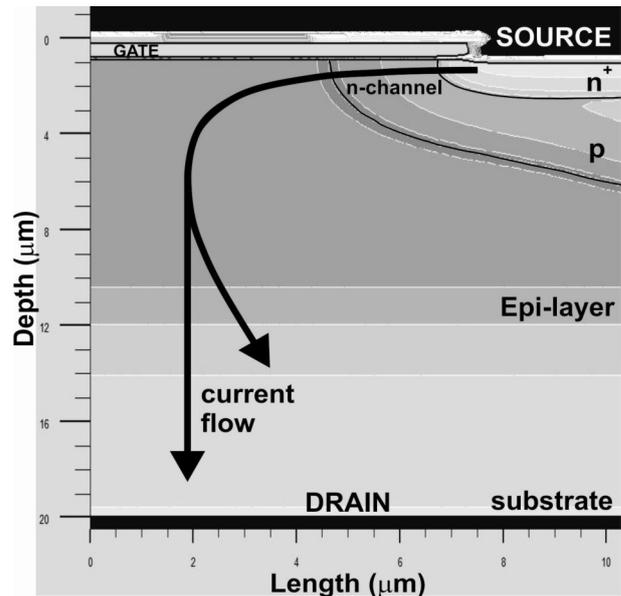
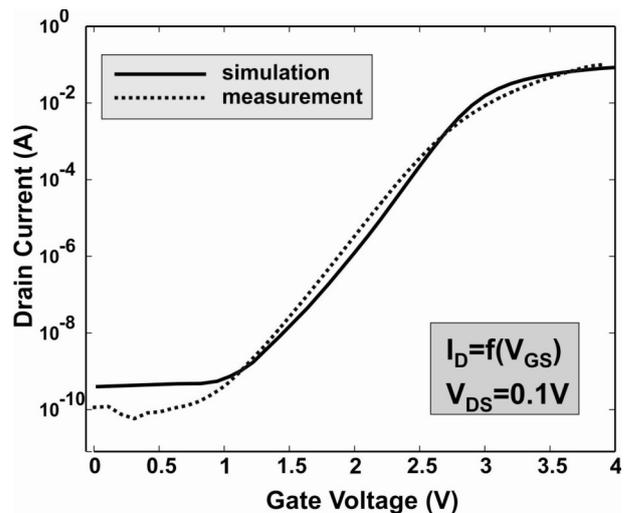**Figure 2:** The cross-section and 2D simulation domain of the power VDMOS transistor.

Figure 3: The measured and simulated transfer characteristics of power VDMOS transistor IRF510.

tation doses and energies, time and temperature of diffusion and oxidation processes, etc., are determined.

The major problem in the described procedure is to fit the gate oxide thickness, channel length and lateral channel doping profile, since they significantly influence on the device electrical characteristics. In addition, because the channel current flows vertically through the n-epitaxial layer and n^+ n^+ substrate to the drain contact (Fig. 2), considerable attention must be given to adjusting the vertical doping profile. Considering all of above data and facts the complete process flow is reconstructed and simulated by using the process simulator ATHENA. The obtained net doping

profile in two-dimensional (2D) simulation domain of power VDMOS transistor is also shown in Fig. 2.

The net doping profile is then used as input parameter for simulation of VDMOS electrical characteristics by using the device simulator ATLAS. Before the electrical simulation, it was necessary to calibrate the parameters of device simulators which largely depend on the specific technology characteristics, such as: fixed oxide charge density at the Si/SiO_2 interface, low-field electron and hole mobility, electron and hole velocity saturation, and electron and hole surface recombination velocities. Finally, a very good agreement of the measured transfer characteristics $I_D = f(V_{GS})$ for $V_{DS} = 0.1V$ and threshold voltage V_{TH} with simulation is obtained, as shown in Fig. 3.

3.2 The influence of semiconductor bulk traps on the electrical characteristics of power VDMOS transistor IRF510

As already mentioned, traps which are generated, for example, due to radiation or the application of high electric field on gate electrode of MOS transistor, can significantly affect on its electrical characteristics. As it is well known, in these cases the defects or traps are formed at the Si/SiO_2 interface, as well as in the oxide and semiconductor bulk.
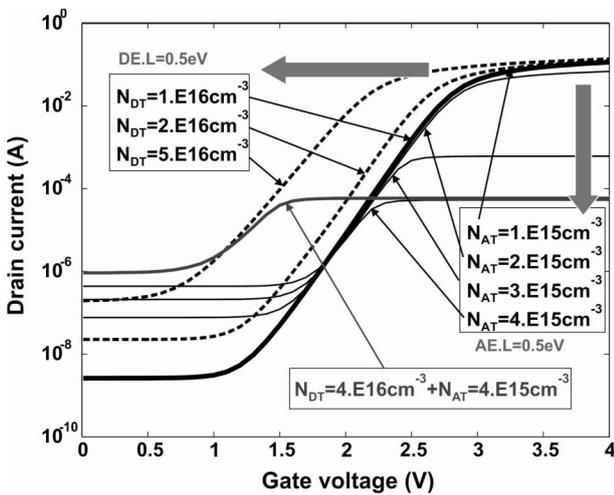

Figure 4: The influence of AT and DT density changes for fixed trap discrete energy levels (*DE.L*=0.5eV, *AE.L*=0.5eV) on power VDMOS IRF510 transfer characteristics.

The subjects of intensive research are the mechanisms of traps generation and the determination of their densities and charges (positive, negative or neutral) [8]. A particular problem is the separation of the effects of different types (DT or AT) and kinds (interface, oxide bulk or semiconductor bulk traps) of defects on the electrical characteristics,

when a well-known experimental methods is used [9, 10]. This problem can be solved efficiently by using the existing TCAD software package tools, where it is possible to separate the influences of different types of defects on the device electrical characteristics. It is also possible to analyze the influence of generated defects on the potential, current density, recombination rate and carriers mobility distribution in the simulation domain.

At the beginning of TCAD analysis of semiconductor bulk traps influence on the electrical characteristics of power VDMOS transistor, it is important to recognize the role of DT and AT on threshold voltage, transconductance, leakage current, transfer and output *I/V* characteristics. In common MOS transistor, drain current flows only through the channel region and therefore only DT or AT influence on the electrical characteristics, depending on whether it is the n-channel or p-channel MOS transistor. In n-channel VDMOS transistors, as is noted above, the drain current after the channel, flows vertically through the n-epitaxial layer and n+ substrate. In this case, DT has the dominant influence on the electrical characteristics in channel region, while AT reduces drain current when it flows vertically through the n epitaxial layer. To verify this assumption, we simulated the electrical characteristics of VDMOS transistor, firstly with DT, for different values of its densities N_{DT}, than with AT, for different values of its densities N_{AT}, and finally when both trap types are presented in semiconductor bulk. The obtained simulation results for given values of N_{DT} and N_{AT}, for fixed energy trap level values *DE.L=0.5eV* and *AE.L=0.5eV* are shown in Fig. 4. Based on the obtained electrical characteristics, it is obvious that when N_{DT} increases, the threshold voltage of VDMOS transistor decreases, while increasing of N_{AT} dramatically reduces the drain current. The leakage current increases in both cases.
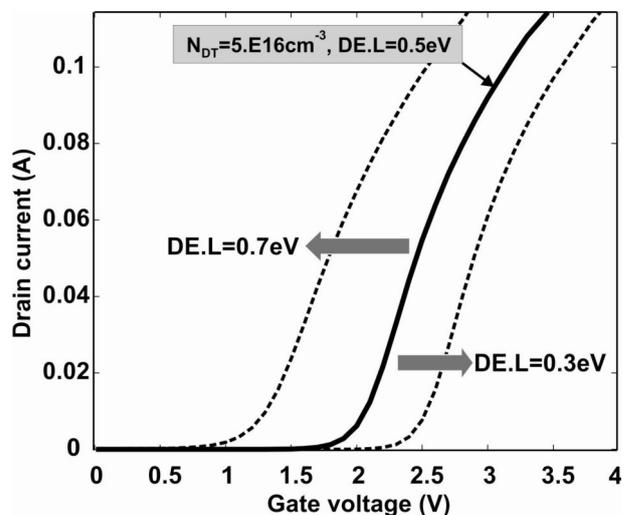

Figure 5: The influence of DT discrete energy level changes for $N_{DT} = 5.10^{16}$ cm⁻³ on power VDMOS transfer characteristics.
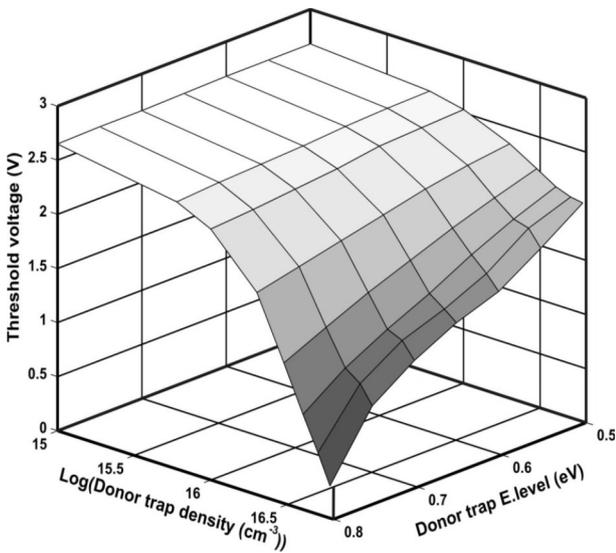
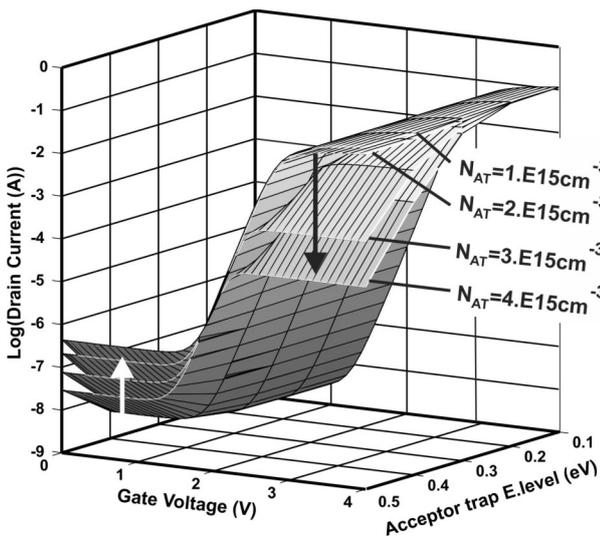**Figure 6:** The changes of the threshold voltage of power VDMOS transistor.

Figure 7: The influence of N_{AT} and *AE.L* on the drain current of power VDMOS transistor.

The influences of AT, when its densities N_{AT} change from 10^{15} to 4.10^{15} cm^{-3}, while the energy level position *AE.L* change from 0.1 to 0.5eV on the drain current are shown in Fig. 7. A dramatic decrease in the drain current with increasing density of AT is obvious. This is the effect of AT influence on the drain current in the n-epitaxial layer as it is discussed earlier. In Fig. 5 the influence of DT discrete energy level changes, for fixed trap densities 4. $N_{DT} = 5.10^{16}$ cm^{-3} is presented. It is obvious that when the value of parameter *DE.L* increases, which means that the donor-like trap energy level is closer to the bottom of the conduction band (Fig. 1.), threshold voltage decreases and at the same time the drain current increases. Vice versa, the threshold voltage increases, when the donor-like trap energy level is

closer to the top of valence band. The changes of the threshold voltage, when the densities of N_{DT} are change in the range from 10^{15} to 10^{17} cm^{-3}, while the energy levels change from 0.5 to 0.8eV, are shown in Fig. 6.

4 Conclusion

In the design of integrated circuits, it is very important to know how the electrical characteristics of individual components are changed under the influences of various stresses. Defects which are generated at the Si/SiO$_2$ interface and in the semiconductor bulk and gate oxide have the dominant influence on the threshold voltage, drain current, transconductance and leakage current of MOS transistor. In this paper we simulate the impact of semiconductor bulk traps, which influences are commonly ignored in standard MOS structures, but in this case, due to the specific structure of the power n-channel VDMOS transistors, where the drain current flows vertically through n-epitaxial layer and substrate, it must be taken into account. Taking the advantage of TCAD software package, the impact of DT and AT influence on the electrical characteristics is simulated and analyzed separately, which is impossible to do with experimental results, where the influences of different mechanisms are mixed. Of course, a complete analysis of HEFS or radiation effect requires the simulation of the impact of the traps generated in gate oxide and at the Si/SiO$_2$ interface on the electrical characteristics of the power VDMOS transistor, which will be a matter of future work.

Acknowledgement

References

1. Anghel C High Voltage Devices for Standard MOS Technologies – Characterisation and Modelling, Ph.D. Thesis No. 3116, Swiss Federal Institute of Technology, Lausanne (EPFL), Lausanne 2004.
2. Wang T, Chiang L, Zous N, Chang T, Huang C Characterization of Various Stress-Induced Oxide Traps in MOSFET's by Using a Subthreshold Transient Current Technique, IEEE Transaction on Electron Devices 1998; 45(8), 1791-1796.
3. Stojadinović N, Danković D, Manić I, Prijić A, Davidović V, Djorić-Veljković S, Golubović S, Prijić

Z Threshold voltage instabilities in p-channel power VDMOSFETs under pulsed NBT stress. Microelectron Reliab 2010; 50(9-11):1278-1282.

4. Alwan M, Beydoun B, Ketata K and Zoaeter M Bias temperature instability from gate charge characteristics investigations in n-channel power MOSFET, Microelectron. J 2007; 38, 727-734.

5. Karim N-M, Soin N, Banitorfian F, Manzoor S Analysis from the perspective of gate material under the effect of negative bias temperature instability, Informacije MIDEM 2012; 42(3): 176-184.

6. Benlatreche M.S, Rahmoune F, Toumiat O Experimental investigation of Si-SiO2 interface traps using equilibrium voltage step technique, Informacije MIDEM 2011; 41(3): 168-170.

7. Cijan G, Tuma T, Tomažič S, Birmen A Fast MOS transistor mismach optimization-a comparison between different approaches, Informacije MIDEM 2009; 39(1): 1-6.

8. Esseni D, Bude J-D, Selmi L On Interface and Oxide Degradation in VLSI MOSFETs—Part II: Fowler–Nordheim Stress Regime, IEEE Transaction on Electron Devices 2002; 49(2), 254-263.

9. Coelho A.V.P, Adam M.C, Boudinov H Distinguishing bulk traps and interface states in deep-level transient spectroscopy, Electroc. Solid State Lett. 2011; 14(9), 368-371.

10. Cartier E Characterization of the hot-electron-induced degradation in thin SiO_2 gate oxide, Microelectron. Reliab. 1997; 38(2), 201-211.

11. Xu H.P.E, Trescases O. P, Sun I.-S. M, Lee D, Ng W.T, Fukumoto K, Ishikawa A, Furukawa Y, Imai H, Naito T, Sato N, Tamura S, Takasuka K, Kohno T Design of a rugged 60 V VDMOS transistor, IET Circuits Devices Syst 2007; 1(5), 327–331.

12. Houssa M, Autran J L, Heyns M M and Stesmans A Model for defect generation at the (1 0 0) SiO2/Si interface during electron injection in MOS structures, Appl. Surf. Sci. 2003; 52, 212–213.

13. Zoaeter M, Beydoun B, Hajjar M, Debs M and Charles J-P Analysis and simulation of functional stress degradation on VDMOS power transistors, Act. Passive Electron Compon 2002; 25(3), 215-238.

14. Grasser T, Selberherr S Modeling of negative bias temperature instability, Invited paper, J. Telecom. and Inform. Technol 2007; 2, 92-102.

15. Aleksić S, Bjelopavlić D, Pantić D Simulation of bulk traps influences on the eleatrical characteristics of VDMOS transistors, In: Proc. TELSIKS 2011 Conf. 2011; 271-275.

16. ATHENA User's Manual, SILVACO, Inc., CA, USA, 2010.

17. ATLAS User's Manual, SILVACO, Inc., CA, USA, 2010.

18. Simmons J. G, Taylor G. W Nonequilibrium steady-state statistics and associated effects for insulators and semiconductors containing an arbitrary distribution of traps, Journal: Physical Review B 1971; 4(2), 502-511.

19. Law M, Solley E, Liang M, Burk D Self-consistent model of minority-carrier lifetime, diffusion length and mobility, IEEE Electron Device Letters 1991; 12(8), 401-403.

20. http://www.vishay.com/docs/91015/sihf510.pdf

Transforming the LSTM training algorithm for efficient FPGA-based adaptive control of nonlinear dynamic systems

Rok Tavčar[1], Jože Dedič[1,2], Drago Bokal[1,3], Andrej Žemva[4]

[1]Cosylab d.d., Control Systems Laboratory, Ljubljana, Slovenia
[2]CO BIK, Solkan, Slovenia
[3]University of Maribor, Faculty of Natural Sciences and Mathematics
[4]University of Ljubljana, Faculty of electrical engineering and CO Namaste,
Ljubljana, Slovenia

Abstract: In the absence of high-fidelity analytical descriptions of a given system to be modeled, designers of model-driven control systems rely on empirical nonlinear modeling methods such as neural networks. The particularly challenging task of modeling time-varying nonlinear dynamic systems requires from the modeling technique to capture complex internal system dynamics, dependent of long input histories. Traditional recurrent neural networks (RNNs) can in principle satisfy these demands, but have limitations on retaining long-term input data. Long Short-Term Memory (LSTM) neural networks overcome these limitations.
In applications with strict requirements imposed on the size, power consumption and speed, embedded implementations of control systems based on Field Programmable Gate Array (FPGA) technology are required. However, as neural networks are traditionally a software discipline, direct ports of neural networks and their learning algorithms into hardware give disappointing, often impractical results. To enable efficient hardware implementation of LSTM with on-chip learning, we present a transformation strategy which leads to replacing original LSTM learning algorithm with Simultaneous Perturbation Stochastic Approximation (SPSA). Our experimental results on a protein sequence classification benchmark confirm the efficacy of the presented learning scheme. The use of this scheme streamlines the architecture of on-chip learning phase substantially and enables efficient implementation of both forward phase and learning phase in FPGA based hardware.

Key words: model predictive control, control of nonlinear dynamic systems, recurrent neural networks, hardware neural networks, FPGA, LSTM, SPSA

Prilagoditev učenja nevronskih mrež LSTM za učinkovito realizacijo adaptivne regulacije nelinearnih dinamičnih sistemov v vezjih FPGA

Povzetek: V primerih kjer podroben analitični opis modela ni na voljo, snovalci modelno naravnanih regulacijskih sistemov potrebujejo empirične nelinearne metode modeliranja kot so umetne nevronske mreže. Modeliranje časovno spremenljivih, nelinearnih dinamičnih sistemov zahteva sposobnost posnemanja zapletene notranje dinamike procesa, pri čemer so izhodi modela odvisni od zgodovine vhodnih podatkov, raztezajoče se prek dolgih časovnih intervalov. Tradicionalne rekurentne nevronske mreže (ang. recurrent neural nework) v principu zadostijo tem zahtevam, ampak imajo omejitve pri pomnjenju vhodov preko dolgih zakasnitev. Posebej z namenom premagati te omejitve so bile zasnovane mreže z dolgim kratkoročnim spominom (ang. Long Short-Term Memory, LSTM).
Mnoge aplikacije, ki imajo stroge zahteve po velikosti, hitrosti in porabi energije, zahtevajo namensko strojno izvedbo regulacijskega algoritma v polju programirljivih logičnih vrat (ang. Field Programmable Gate Array, FPGA). Ker so nevronske mreže tradicionalno disciplina splošnonamenske programske opreme, neposredna preslikava nevronskih mrež in njihovega algoritma za učenje v strojno opremo običajno prinese nepraktičen rezultat z zmogljivostmi pod pričakovanji. Z namenom učinkovite realizacije mrež LSTM z učenjem v strojni opremi, v tem delu predstavljamo prilagoditveno strategijo, ki motivira zamenjavo izvirnega učnega algoritma z algoritmom Simultaneous Perturbation Stochastic Approximation (SPSA). Učinkovitost delovanja mrež LSTM, učenih z SPSA, potrdimo s poskusi na znanem učnem problemu klasifikacije beljakovin. Nova kombinacija arhitekture nevronske mreže ter algoritma za učenje omogoča izjemne poenostavitve pri izvedbi tako testne faze kot učne faze v namenski strojni opremi, osnovani na tehnologiji FPGA.

Ključne besede: prediktivno vodenje, vodenje nelinearnih dinamičnih sistemov, rekurentne nevronske mreže, nevronske mreže v strojni opremi, FPGA, LSTM, SPSA

* Corresponding Author's e-mail: rok.tavcar@cosylab.com

1 Introduction

Control of complex nonlinear dynamical systems demands advanced control methods such as Model Predictive Control (MPC). MPC is particularly challenging in the absence of a high-fidelity analytical description of the modeled system, a frequent reality in control of real-world systems. In such cases, designers must rely on empirical nonlinear modeling methods such as neural networks (NN) [1]. Neural-network-based modeling also plays an important role in control of time-varying systems, where NN learning is used to adapt to system-parameter changes over time [2]. Typical MPC tasks demand the model to capture complex internal system dynamics, dependent on long input histories. The type of neural networks that can satisfy these demands are Recurrent Neural Networks (RNNs) [3]. Because of their enhanced prediction qualities, they have been applied to numerous dynamic system control applications including speech recognition, phoneme recognition and chemical process identification [4]. However, traditional RNN models have limitations on retaining long-term input data. Long Short-Term Memory (LSTM) neural networks have been designed particularly to overcome these limitations by introducing architectural concepts which prevent exponential decay of input information over extended sequence lengths [5]. These concepts enable LSTM networks to learn patterns in sequences longer than 1000 steps, a 2 orders of magnitude improvement over traditional RNNs [6]. Figure 1 shows a basic unit of an LSTM network called a memory block.
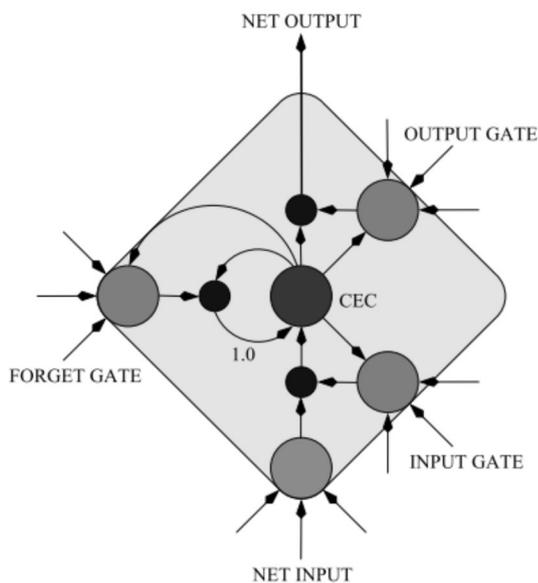

Figure 1: LSTM Memory Block: the memory cell with its constant error carousel (CEC) retains data over long input sequences. The update, output and erasure of this data are controlled by input, output and forget gates, respectively. Image source: [7].

It comprises several neuron-like components (a memory cell 'guarded' by gating units; input, output and forget gates), each with its own set of in- and outcoming weighted connections and a nonlinear activation function.

In control applications with strict requirements imposed on size, power consumption and speed, compact implementations of control systems in dedicated hardware are required. Due to the ceaselessly increasing density of Field Programmable Gate Arrays (FPGAs), along with their high degree of flexibility, they have become the technology of choice in a wide range of control applications [8]. However, NNs being traditionally a software discipline, direct ports of themselves and their learning algorithms into hardware give disappointing, often impractical results. Thus, algorithmic and architectural transformation techniques are crucial when porting neural networks into hardware [9].

In this work, we aim towards hardware-friendly implementation of LSTM with on-chip learning. This paper presents a strategy by which LSTM training is transformed and adapted in a way that reduces overall architectural complexity and allows hardware implementation with explicit exploitation of parallelism, avoiding mechanisms requiring complex circuitry.

Moreover, our proposed implementation strategy enables an independent architectural design of network forward phase and learning phase, leaving wide design freedom in choosing the implementation approach for each phase.

The validity of our approach is confirmed by experiments presented in this paper, showing that our proposed approach, i.e. learning of LSTM with Simultaneous Perturbation Stochastic Approximation (LSTM-SPSA), retains the ability of LSTM to learn sequences in data whilst delivering immense architectural benefits in terms of suitability for hardware implementation.

The paper is organized as follows. The rest of Chapter 1 briefs our mission statement and reviews related work. Chapter 2 explains the proposed transformation strategy and motivates the search of an alternative algorithm for LSTM learning. This search is laid out in Chapter 3, which also explains the chosen algorithm and emphasizes the advantages and drawbacks of the new learning scheme. Chapter 4 explains and discusses our experiments and results. Chapter 5 provides the conclusion and guidelines for future work.

1.1 Mission statement

In this work, we make the first attempt to transform LSTM and its learning rule to enable their efficient im-

plementation in dedicated hardware. At the time of writing, no account on research aiming at a hardware-native implementation of LSTM and its learning rule has yet been published.

We seek for the optimal strategy for efficient hardware implementation of both LSTM forward pass and on-chip learning.

We stress the importance of early architectural transformations upon porting software algorithms into dedicated hardware. We are led by the idea that an early, educated algorithm transformation will yield superior gains compared to a low-level, partial optimization of a design based on concepts unfit for dedicated hardware.

Investing effort to review alternative implementation options is crucial ground work that enables early architectural decisions that maximize future design fitness.

1.2 Related work

In the last two decades, extensive experimental and theoretical research effort has been aimed towards optimal hardware realization of different NN types and learning algorithms [9,10,11,12].

Systolic arrays [13, ch.5], [14] and stream processing architectures [15] minimize hardware idle-time and optimize dataflow and hardware efficiency. However, these approaches put strong constraints on the kind of neural architectures they can be applied to [13]. Logarithmic multipliers [16,17] spare hardware resources needed to implement expensive multipliers. Such optimization techniques gain priority when the benefits of higher-level algorithmic transformations have already been exploited. Limiting network weight levels to powers of two [18] replaces multipliers altogether with bitwise shifts. However, typical neural networks do not allow such modifications without significant performance loss. Cellular neural networks [19] and RAM-based NNs [9] are specifically designed with efficient hardware implementation in mind. However, their implementation principles (and their learning algorithms, also typically suitable for hardware) cannot be arbitrarily transferred to other network architectures, rendering these implementation principles unsuitable for applications requiring specific architectures. Perturbation algorithms and local learning algorithms [9] generalize well to different network architectures, and are well-suitable for hardware implementations. Perturbation algorithms do not put any assumptions on the neural network architecture, which is particularly beneficial when they are applied to architecturally complex neural networks such as LSTM.

Specifically for LSTM networks, no development on their architecture or their learning algorithm has yet been aimed at improving their suitability for hardware implementation. Improvements of LSTM are mainly focused towards improving their learning capacity [20,21] or convergence [22]. Research has yet to be made towards making LSTM networks and their learning suitable for dedicated hardware.

2 Criteria for selecting the transformation approach

In our search for conceptual transformations to LSTM and its learning on the algorithmic level, alternatives that bring the following benefits are sought for:

- decoupling the implementation of forward and backward pass to reduce implementation complexity, possibly without doubling the necessary hardware resources
- lowering the amount of expensive arithmetic operations
- lowering the complexity of control circuitry
- lowering the data dependencies between different algorithm parts (improving spatial locality)

To keep complexity of the hardware implementation at minimum, the implementation of on-chip learning should affect the implementation of the network's forward phase as little as possible. Ideally, the two phases should be completely decoupled, the only link between them being the data they both need in their operations (e.g. they both access the same memory storing network weights). In such case, the design of each phase can be treated separately, giving the designer more flexibility when choosing design approaches for either of them. However, being based on backpropagation, the LSTM backward pass is in the same order of architectural complexity as the forward pass, thus complete separation of the two phases could mean doubling the amount of required hardware resources. This motivates an architectural design where parts of the hardware are used by both phases; but that complicates the implementation process significantly compared to the case where each phase is treated independently. Consequently, we seek high-level transformations that allow the design of backward pass independently from the forward pass without a significant increase of the overall required hardware resources.

As the first step, we systematically analyze the findings of our literature search laid out in the previous chapter with respect to our criteria. As LSTM's advanced learning abilities stem from its architectural composition, we leave the neural network topology intact and focus on

augmenting the LSTM learning rule. We isolate hardware-friendly learning algorithms that generalize well to different neural network topologies and satisfy our criteria in several points. In subsequent steps of our research, these algorithms are analyzed in further depth.

3 Selecting the alternative training algorithm for LSTM

There are in principle two classes of hardware-friendly training algorithms: a) variations of widely-used but complex training algorithms with some of their core mechanisms altered or replaced and b) training algorithms that apply to hardware-friendly network architectures and are thus, in concept, fit for hardware themselves [11].

Because LSTM networks are a traditional multilayer network architecture and original LSTM training is based on backpropagation, it is best to look for algorithms close to its principles, focusing thus on the first class of learning algorithms. Their most successful representatives rely on some variety of parameter perturbation.

The general idea of perturbation algorithms is to obtain a direct estimate of the gradients by a slight random perturbation of network parameters, using the forward pass to measure the resulting network error. These on-chip training techniques do not only eliminate the complex backward pass but are also likely to be more robust to non-idealities occurring in hardware, such as a lowered numerical precision [9]. Mainly two variations exist: node perturbation and weight perturbation. Examples of node perturbation algorithms are Madaline-3 and Madaline-2.

We choose weight perturbation algorithms because of the lower complexity of their addressing and routing circuitry compared to node perturbation algorithms. Specifically, we look into two fully parallel versions of weight perturbation algorithms, namely Simultaneous Perturbation Stochastic Approximation (SPSA) [23] and Alopex [24]. Both are local training algorithms which determine weight updates using only locally available information and a global error signal. Both algorithms are closely related, but unlike SPSA, Alopex relies on the principles of simulated annealing, which adds complexity to the calculation of each weight perturbation.

In contrast, SPSA uses a simple random distribution function to perform weight perturbations and then updates all weights using the same absolute value of the update. Neither algorithm makes any assumptions as to the neural network topology, thus both are

conceptually fit for direct generalization to LSTM network architecture. Neither have yet been applied to the LSTM architecture, but have been demonstrated to successfully train simpler FFNNs and RNNs [24, 25, 26], which motivates us to research their applicability for LSTM training. Because SPSA uses less parameters and computational steps to determine the update of each weight than Alopex, ultimately allowing a more streamlined hardware description, SPSA was selected as the algorithm of choice in this study.

3.1 LSTM-SPSA: LSTM trained by Simultaneous Perturbation Stochastic Approximation

SPSA [23] is based on a low-complexity, highly efficient gradient approximation that relies on measurements of the objective function, not on the gradient itself. The gradient approximation is based on only two function measurements, regardless of the dimension of the gradient vector, which is especially important in the field of neural networks, where this dimension quickly reaches several thousands. The weight-update scheme of the SPSA learning algorithm is explained by the following equations:

$$\Delta w_t^i = \frac{J(w_t + cs_t) - J(w_t)}{cs_t^i} \tag{1}$$

$$\Delta w_{t+1}^i = \begin{cases} w_{\max} & , if\,(w_t^i - a\Delta w_t^i) > w_{\max} \\ -w_{\max} & , if\,(w_t^i - a\Delta w_t^i) < -w_{\max} \\ w_t^i - a\Delta w_t^i, & otherwise \end{cases} \tag{2}$$

Here Δw_t and Δw_t^i denote the weight vector of a network and its i-th element at the t-th iteration, respectively, α is a positive constant and c is the magnitude of the perturbation. Δw^i represents the i-th element of the modifying vector. $W_{\max}$ is the maximum value of a weight. s_t and s_t^i denote a sign vector and its i-th element that is 1 or -1. The sign of s_t^i is determined randomly, with adherence to one of the recommended variable distributions. $J(w_t)$ denotes the criterion function, which is most frequently the Mean Square Error (MSE) between the network's actual and desired output.

From Eqs. 1 and 2 we see that a) during weight update, the *same absolute value* is used to update *all* network weights and b) to compute this value, only *two measurements* of the error function are required, one obtained via forward pass with perturbed weights and one without perturbations. SPSA algorithm flowchart is shown in Figure 2.
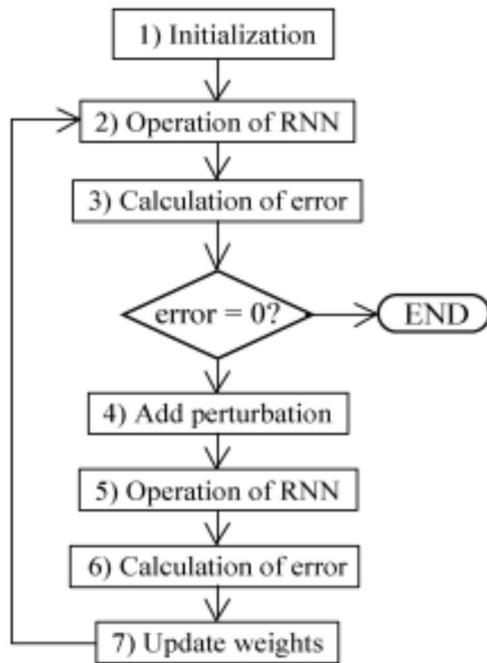
Figure 2: Flowchart of Simultaneous Perturbation Stochastic Approximation applied to Recurrent Neural Network Learning. Image source: [26].

The first advantage of SPSA over the original LSTM learning algorithm is simplification of the gradient estimation, because of the substantial reduction the number of arithmetical operations needed for weight updates. The second advantage, less obvious but equally important, is SPSA's equal treatment of all weights, eliminating in this way the separate error backpropagation paths (with different arithmetic expressions) required by different LSTM weight types, simplifying the algorithm routing circuitry significantly.

This second advantage in simplicity could prove to be a disadvantage in learning performance. For example, error backpropagation paths (set of weighted connections) that lead into forget gates, could have entirely different update dynamics than those leading into input gates. In original LSTM learning, this is accounted for; but not in SPSA. It is thus expected that SPSA algorithm will take longer to converge than original LSTM learning rule, but the increased simplicity of hardware implementation could compensate this by increasing operation speeds and possibilities of parallelization. An added benefit is also a simpler, more easily maintainable hardware description code.

3.2 Improving Learning Performance of LSTM-SPSA

After initial experiments with LSTM-SPSA (on the benchmark presented in the following chapter), possible augmentations to the learning algorithm were ex-

plored to maximize learning performance. The underlying idea of our augmentation was that if presented with a more difficult task, the algorithm will also improve on its basic task (minimize mean square error). For classification tasks such as ours, receiver operating characteristics curves (ROC) are better discriminative measures of classification performance than MSE [28]. Furthermore, the classification performance is in our experiments measured by AUC and AUC50 (area under ROC and ROC50 curve, respectively, presented briefly in the next chapter), [6], motivating the idea that the algorithm should also aim to maximize these scores.

To challenge our learning algorithm with a more difficult optimization task, we extended the criterion function by adding the AUC and AUC50 score, getting two new criterion functions. In addition to bringing MSE towards zero, the algorithm thus also had to maximize (bring to value of 1) AUC or AUC50. The two enhanced criterion functions used were:

$$J_{AUC}(w_t) = MSE + y * (1 - AUC) \text{ and}$$

$$J_{AUC50}(w_t) = MSE + y * (1 - AUC50)$$

using y as a scaling factor to tune the ratio between the MSE and AUC (or AUC50) in the score.

Because the AUC score can only be calculated at the end of a learning epoch, we needed to implement batch learning, applying cumulative weight updates at each learning epoch end. When using batch learning with the original criterion function, the performance of the learning algorithm did not change significantly compared to online learning. When adding ROC or ROC50 momentum to the criterion function, learning improved only by a few %, not reaching statistical significance.

4 Experimental results

Replacing the learning algorithm considerably interferes with the neural network's native learning scheme. Thus, before actual hardware implementation, the effectiveness of SPSA in training of LSTM has to be experimentally verified.

The most significant property of LSTM networks is their ability to retain temporal information in sequential input data. Therefore, we must test the LSTM-SPSA combination on a learning task that demands this ability. To allow for a back-to-back comparison with the original implementation, our experiments were based on those described in [6]. We implemented SPSA learning for LSTM networks and applied LSTM-SPSA to the

SCOP 1.53 database, which is a standard, widely used sequence-classification benchmark.

The preliminary experiments on a single SCOP 1.53 dataset, described in [27], showed promising learning results, indicating that SPSA-trained LSTM networks are able to learn temporal information over extended lengths of sequences.

For the main experiment, run on the complete SCOP 1.53 benchmark, we used pure SPSA with the original criterion function on an LSTM NN architecture identical to the one described in [6]. We used online learning, meaning that weight updates were computed and applied at the end of each sequence presented to the network within a learning epoch. In the experiment, the two SPSA learning parameters values used were c=0.0015 and $a = \dfrac{c}{2^3}$. In the generation of SPSA perturbation matrix, a Bernoulli distribution was used, as one of the recommended, optimal distributions for SPSA perturbations [29].

Table 1 shows the performance of different algorithms applied to SCOP 1.53 benchmark, showing that LSTM NNs outperform traditional algorithms for protein sequence classification in terms of classification quality, speed or both [6]. The quality of a ranking of test set examples for each protein family is evaluated by the area under the ROC curve. Being a more discriminiative quality measure, the area under ROC50 is also used; this is the area under the ROC curve up to 50 false positives, essentially rescaling the false positive rate of the ROC curve [6]. ROC and ROC50 scores for LSTM-SPSA show competitive learning performance of LSTM-SPSA towards other protein sequence classification algorithms. Because the forward phases of LSTM and LSTM-SPSA are identical, their test times, (Table 1, column 3) when run on software, are equal. Results in the table confirm that after replacing the original LSTM learning algorithm with SPSA, the learning ability of the LSTM NN architecture is preserved to a high degree. Because of the computational and architectural advantages of SPSA, explained in chapter 3, this motivates the use of LSTM-SPSA in hardware implementations of solutions that require the unique learning abilities of LSTM NN architecture.

Table 1: Results of remote homology detection on the SCOP benchmark database. The second and third column report the average area under the receiver operating curve ('ROC') and the same value for maximally 50 false positives ('ROC50'). The fourth column reports the time required to classify 20 000 test protein sequences (equivalent to one genome) into one superfamily. Performance data for solutions other than LSTM-SPSA sourced from [6].

Method	ROC	ROC50	Time
PSI-BLAST	0.693	0.264	5.5 s
FPS	0.596	-	6800 s
SAM-T98	0.674	0.374	200 s
Fisher	0.887	0.250	> 200 s
Mismatch	0.872	0.400	380 s
Pairwise	0.896	0.464	> 700 s
SW	0.916	0.585	> 470 s
LA	0.923	0.661	550 h
Oligomer	0.919	0.508	2000 s
HMMSTR	-	0.640	> 500 h
Mismatch-PSSM	0.980	0.794	> 500 h
SW-PSSM	0.982	0.904	> 620 h
LSTM	0.932	0.652	20 s
LSTM-SPSA	0.900	0.392	20 s

Figure 3 and Figure 4 show the total number of families for which a given algorithm exceeds a ROC or ROC50 threshold, respectively. Because of the rescaling of false positives in ROC50 score, giving it a higher discriminative value, the difference in performance between LSTM and LSTM-SPSA is more evident in Figure 4.
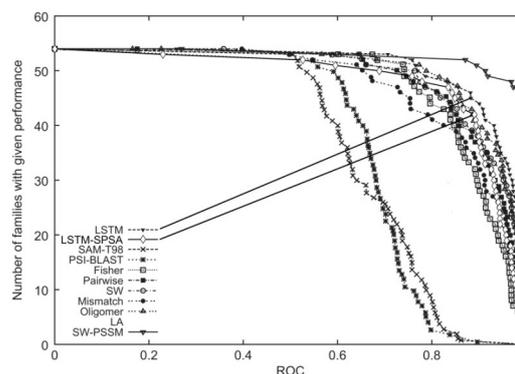

Figure 3: Comparison of homology detection methods for the SCOP 1.53 benchmark dataset. The total number of families for which a given method exceeds a ROC threshold is plotted. Performance data for solutions other than LSTM-SPSA sourced from [6].

Performance figures show that LSTM-SPSA exhibits competitive results compared to other protein classification techniques and compares to the original learning algorithm. This confirms that LSTM-SPSA retains the ability of LSTM networks to learn long sequences in data and, due to its substantial architectural advantages, that it is a viable scheme for implementing LSTM network abilities in dedicated hardware.

5 Conclusion

The work presented in this paper is the first attempt in transforming LSTM and its learning rule with the aim of
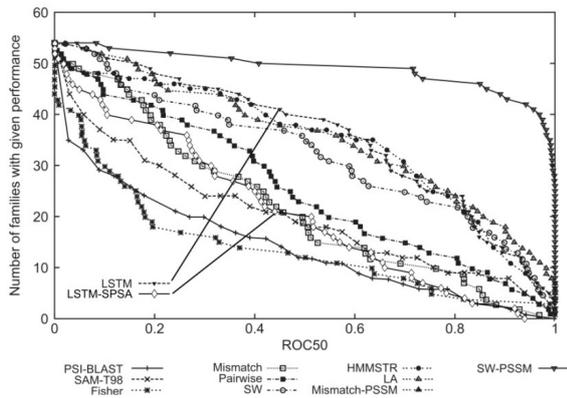
Figure 4: Comparison of homology detection methods for the SCOP 1.53 benchmark dataset. The total number of families for which a given method exceeds a ROC50 threshold is plotted. Performance data for solutions other than LSTM-SPSA sourced from [6].

improving its suitability for hardware implementation. Our transformation strategy is based on the premise that most gains can be achieved by high-level transformations of the algorithm on a conceptual level, which can mean completely replacing its vital parts with alternatives known to be suitable for hardware.

In our particular case, we have refrained from a naive direct port of a LSTM learning algorithm from software to hardware platform, bound to give disappointing results. Instead, we have replaced LSTM's backpropagation-based learning algorithm with Simultaneous Perturbation Stochastic Approximation, which fits our criteria for suitability for hardware implementation.

Our experiments confirm that LSTM-SPSA retains its ability to learn patterns in sequential data, which is the main characteristic of the LSTM network architecture. Due to promising results on a classification task, we expect that LSTM-SPSA could also demonstrate regression abilities. Our results show that LSTM-SPSA yields competitive results to the original learning algorithm, while enabling a cleaner implementation, lower resource utilization, simpler logical circuitry and increased parallelization of LSTM with on-chip learning.

Our strategy yields a solution which enables the designer to treat the forward phase and learning phase circuitry separately and to seek implementation strategies for each independently, giving a broader set of possibilities. Moreover, as SPSA is significantly less complex than the original algorithm, this decoupling does not bring a large increase of FPGA fabric consumption.

We conclude that because of the ability of SPSA in training LSTM on sequential data and because of its substantial advantages in suitability for hardware im-

plementation, LSTM-SPSA is the recommended approach for dedicated hardware implementations of LSTM networks with on-chip learning.

In our future work, the effects of precision loss due to fixed-point arithmetic used in hardware will be studied. Preliminary experiments show that different fixed-point scaling should be used for different parts of the NN. Regression abilities of LSTM-SPSA will be explored. An attempt will be made to improve LSTM-SPSA learning either by using a modified SPSA which uses smoothed gradient or by using an adaptive learning rate. Independently from the learning phase, transformation techniques for LSTM forward phase will be reviewed.

Acknowledgements

References

1   T. Hayakawa, W. M. Haddad and N. Hovakimyan, "Neural network adaptive control for a class of nonlinear uncertain dynamical systems with asymptotic stability guarantees", Neural Networks, IEEE Transactions on, vol.19, no.1, pp. 80-89 2008.

2   H. Chaoui, P. Sicard and W. Gueaieb, "Ann-based adaptive control of robotic manipulators with friction and joint elasticity", Industrial Electronics, IEEE Transactions on, vol.56, no.8, pp. 3174-3187 2009.

3   R. K. Al Seyab and Y. Cao, "Nonlinear system identification for predictive control using continuous time recurrent neural networks and automatic differentiation", Journal of Process Control, vol.18, no.6, pp. 568-581 2008.

4   P. A. Mastorocostas and J. B. Theocharis, "A recurrent fuzzy-neural model for dynamic system identification", Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, vol.32, no.2, pp. 176-190 2002.

5   S. Hochreiter and J. Schmidhuber, "Long short-term memory", Neural Computation, vol.9, no.8, pp. 1735-1780, 1997/11/01 1997.

6   S. Hochreiter, M. Heusel and K. Obermayer, "Fast model-based protein homology detection without alignment", Bioinformatics, vol.23, no.14, pp. 1728-1736, July 15, 2007 2007.

7   A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures", Neural Networks, vol.18, no.5-6, pp. 602-610 2005.

8   E. Monmasson, L. Idkhajine, M. N. Cirstea, I. Bahri, A. Tisan and M. W. Naouar, "Fpgas in industrial control applications", Industrial Informatics, IEEE Transactions on, vol.7, no.2, pp. 224-243 2011.

9   P. Moerland and E. Fiesler, "Neural network adaptations to hardware implementations", in Handbook of Neural Computation, 1997.

10   D. G. Bailey, "Invited paper: Adapting algorithms for hardware implementation", in Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on, pp. 177-184, 2011.

11   P. Moerland and E. Fiesler, "Hardware-friendly learning algorithms for neural networks: An overview", in Proceedings of the Fifth International Conference on Microelectronics for Neural Networks and Fuzzy Systems: MicroNeuro'96, IEEE Computer Society Press, Lausanne, Switzerland, 1996.

12   J. Misra and I. Saha, "Artificial neural networks in hardware: A survey of two decades of progress", Neurocomputing, vol.74, no.1-3, pp. 239-255 2010.

13   A. Omondi and J. C. Rajapakse, Fpga implementations of neural networks, Springer Netherlands, 2006.

14   R. G. Gironés, R. C. Palero, J. C. Boluda and A. S. Cortés, "Fpga implementation of a pipelined on--line backpropagation", The Journal of VLSI Signal Processing, vol.40, no.2, pp. 189-213 2005.

15   C. Farabet, C. Poulet and Y. LeCun, "An fpga-based stream processor for embedded real-time vision with convolutional networks", in Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on, pp. 878-885, 2009.

16   U. Lotrič and P. Bulić, "Logarithmic multiplier in hardware implementation of neural networks", Adaptive and Natural Computing Algorithms, pp. 158-168 2011.

17   U. Lotrič and P. Bulić, "Applicability of approximate multipliers in hardware neural networks", Neurocomputing, vol.96, no.0, pp. 57-65 2012.

18   M. Marchesi, G. Orlandi, F. Piazza and A. Uncini, "Fast neural networks without multipliers", Neural Networks, IEEE Transactions on, vol.4, no.1, pp. 53-62 1993.

19   L. Fortuna, P. Arena, D. Balya and A. Zarandy, "Cellular neural networks: A paradigm for nonlinear spatio-temporal processing", Circuits and Systems Magazine, IEEE, vol.1, no.4, pp. 6-21 2001.

20   F. A. Gers, J. Schmidhuber and F. Cummins, "Learning to forget: Continual prediction with lstm", Neural Computation, vol.12, no.10, pp. 2451-2471, 2000/10/01 2000.

21   F. A. Gers, N. N. Schraudolph and J. Schmidhuber, "Learning precise timing with lstm recurrent networks", Journal of Machine Learning Research, vol.3, pp. 115-143 2002.

22   J. Schmidhuber, D. Wierstra, M. Gagliolo and F. Gomez, "Training recurrent networks by evolino", Neural Computation, vol.19, no.3, pp. 757-779 2007.

23   J. C. Spall, "Implementation of the simultaneous perturbation algorithm for stochastic optimization", Aerospace and Electronic Systems, IEEE Transactions on, vol.34, no.3, pp. 817-823 1998.

24   K. Unnikrishnan and K. Venugopal, "Alopex: A correlation-based learning algorithm for feedforward and recurrent neural networks", Neural Computation, vol.6, no.3, pp. 469-490 1994.

25   Y. Maeda and R. J. P. De Figueiredo, "Learning rules for neuro-controller via simultaneous perturbation", Neural Networks, IEEE Transactions on, vol.8, no.5, pp. 1119-1130 1997.

26   Y. Maeda and M. Wakamura, "Simultaneous perturbation learning rule for recurrent neural networks and its fpga implementation", Neural Networks, IEEE Transactions on, vol.16, no.6, pp. 1664-1672 2005.

27   R. Tavčar, "Design of neural networks for dedicated hardware implementation", in Proceedings of Microelectronics, Devices and Materials (MIDEM), International Conference on, MIDEM, 2012, Otočec, Slovenia, 2012.

28   J. Huang and C. X. Ling, "Using auc and accuracy in evaluating learning algorithms", Knowledge and Data Engineering, IEEE Transactions on, vol.17, no.3, pp. 299-310 2005.

29   P. Sadegh and J. C. Spall, "Optimal random perturbations for stochastic approximation using a simultaneous perturbation gradient approximation", Automatic Control, IEEE Transactions on, vol.43, no.10, pp. 1480-1484 1998.

MIDEM 2013

49th INTERNATIONAL CONFERENCE ON MICROELECTRONICS, DEVICES AND MATERIALS WITH THE WORKSHOP ON DIGITAL ELECTRONIC SYSTEMS

CALL FOR PAPERS

Announcement and Call for Papers
September 25th – 27th, 2013
Hotel Larix, Kranjska Gora, Slovenia

ORGANIZER: MIDEM Society - Society for Microelectronics, Electronic Components and Materials, Ljubljana, Slovenia

CO – ORGANIZER: CO NAMASTE - Centre of Excellence, Ljubljana, Slovenia

CONFERENCE SPONSORS: Slovenian Research Agency, Republic of Slovenia; IMAPS, Slovenia Chapter; IEEE, Slovenia Section; Zavod TC SEMTO, Ljubljana.

GENERAL INFORMATION
The 49th International Conference on Microelectronics, Electronic Components and Devices with Workshop on Digital Electronic Systems continues the tradition of the annual international conferences organised by MIDEM, Society for Microelectronics, Electronic Components and Materials, Ljubljana. The conference will be held in **Kranjska Gora**, Slovenia, well-known ski resort and conference centre, from **SEPTEMBER 25th – 27th, 2013**.

Topics of interest include but are not limited to:
- Novel monolithic and hybrid circuit processing techniques,
- New device and circuit design,
- Process and device modelling,
- Semiconductor physics,
- Sensors and actuators,
- Electromechanical devices, Microsystems and nanosystems,
- Optoelectronics,
- Photovoltaic devices,
- New electronic materials and applications,
- Electronic materials science and technology,
- Materials characterization techniques,
- Reliability and failure analysis,
- Education in microelectronics, devices and materials.

ABSTRACT AND PAPER SUBMISSION:
Prospective authors are cordially invited to submit up to 1 page abstract before **May 1st, 2013**. Please, identify the contact author with complete mailing address, phone and fax numbers and e-mail address.

After notification of acceptance (**June 1st, 2013**), the authors are asked to prepare a full paper version of six pages maximum. Papers should be in black and white. Full paper deadline in PDF and DOC electronic format is: **August 1st, 2013.**

IMPORTANT DATES:
- Abstract deadline: **May 1st, 2013** (1 page abstract or full paper)
- Notification of acceptance: **June 1st, 2013**
- Deadline for final version of manuscript: **August 1st, 2013**

Invited and accepted papers will be published in the conference proceedings.

Deatailed and updated information about MIDEM Conferences is available at:
http://www.midem-drustvo.si/ under Conferences.
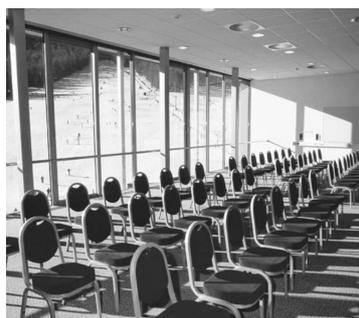
Boards of MIDEM *Society | Organi društva* MIDEM