Silicon Approximation to Biological Neuron

V.A. Gorelik
Neuronix
7003 Sand Road
Savannah, GA 31410

***Abstract* – This paper presents a new approach to simulate the behavior of a biological neuron in silicon. The proposed device has the ability to mimic a variety of structures and interconnect architectures commonly found in biological neural nets. The proposed device may be fabricated in polysilicon rather than in a single crystal substrate and thus permits multi-layer architectures. Unlike MOS and BJT-based structures, the device utilizes significantly different operating principles that are more closely related to transport mechanisms in biological neurons. The device is well suited to simulate axodendritic, dendrodendritic, axoaxonic, and reciprocal synapses. A network of such devices can be constructed to perform both spatial and/or temporal processing. Basic principles underlying the design allow multi-layered, almost zero-power neural networks on a single silicon die. One possible implementation, utilizing temporal neural circuitry for extraction and production of atomic auditory elements – phonemes in Broca's and Wernicke's areas of the cortex, is also shown.**

I. Introduction.

Biological neurons consist of a cell body with dendrites branching-out from it, and an axon emerging from the cell body generally in the opposite direction. The majority of the neuron's surface (cell membrane), except for the sheathed (myelinated) axon, is covered with synaptic sites. Neurons communicate with each other via a variety of synaptic arrangements: Axodendritic (Axon to Dendrite), Dendrodendritic (Dendrite to Dendrite), Axoaxonic (Axon to Axon), and Reciprocal Synapses. The latter is formed when two or more dendrites are juxtaposed to form a synapse for bidirectional communication. More complex arrangements may also exist, involving more than two neurons.

Perception is a dynamic neural process, and thus timing is a key consideration. In 1949 Donald Hebb proposed the theory of temporal control in biological systems based on the concept of reverberating cell assemblies [1]. Recently, much work has been done on reverberating behavior and its impact on mechanisms of perception and cognition [2]. According to William H Calvin [3] "It is important to extend this concept to a level when sequential activation of neural elements can produce thought and action. There is mounting evidence that brains do use population codes that are sensitive to temporal relationships on various time scales in order to exhibit motor behavior, speech, language, vision, audition, and reasoning." A key problem turns out to be to engineer systems that can simulate the spatio-temporal relationship between neurons, and to understand how they collectively encode and decode information.

Properties of biological synapses depend on their size and metabolic rate [4], while training is the process of adjusting these properties to meet a specific goal. In artificial (silicon) neurons [5] the same is sometimes achieved by implementing non-volatile memory [6], based on the physics of floating gate structures as storage elements for holding tunable synaptic weight coefficients. A biological brain can process very complex sensory patterns in both spatial and temporal domains. The brain's responses to environmental changes can range from simple reflexes to generation of abstract thoughts and emotions. Unlike the brain's circuitry, most ANNs (Artificial Neural Networks) can process only low-level spatial stimuli, and provide very limited means for temporal processing in complex hierarchical and multidimensional environments. Some work has been done in the theory of ANNs to introduce temporal behavior [7], [8], however practical implementations of artificial neurons for temporal processing remain challenging. The proposed concept provides a compact and efficient means for a higher degree of spatial and temporal processing within Artificial Neural Networks.

One factor that significantly limits further improvement in structural complexity of ANNs is the lack of adequate technology for interconnects. The complexity of modern ANNs remains orders of magnitude lower than that of biological nets. State of the art approaches allow simulating neural behavior only in terms of axodendritic connections and thus are limited.

An unorthodox electronic component based on the physics of a giant floating gate structure to provide more flexibility in building neural networks for both spatial and temporal processing will be discussed. It allows the implementation of axodendritic, dendrodendritic, axoaxonic, and reciprocal synaptic arrangements resulting in greatly improved interconnect flexibility of corresponding ANNs.

II. General description.

A silicon neuron contains a Giant Polysilicon Floating Structure (GPFS) as the main computational element and is based on the ability of GPFS to preserve electric charge. Mostly impermeable SiO_2 or Si_3N_4 barrier insulates GPFS from the surrounding circuitry and the substrate. In at least two locations (Injection and Tunneling nodes) the thickness of this barrier is reduced to make the injection and removal of electrical charges (in this case electrons) possible.

GPFS is made from undoped or low-doped polysilicon to allow an electic field to exist in its bulk. As a result, the local charge density within the GPFS' volume becomes a function of externally applied electric fields. The body of the GPFS branches-out (similar to neuron's dendrites) and these branches are the basis for artificial synapses.

The device performs summation of a large number of weighted and slow-changing input stimuli, which are presented in analog, or pulse-width modulated form. The output can easily be converted into either analog value or a train of width modulated pulses.

The proposed model of the silicon neuron utilizes the ability of a GPFS to store an electric charge and perform signal processing by forcing the redistribution of this charge in the GPFS by the means of external electric fields. A charge is normally injected into and tunneled out of the GPFS by hot electron injection and Fouler-Nordheim tunneling mechanisms. Electric fields are applied to the GPFS via capacitively coupled polysilicon or metal electrodes. These electrodes are located below and/or above the GPFS. They also can be fabricated within the substrate in a form of heavily doped n+ or p+ areas or a CCD channel, etc. These externally applied electric fields determine the energy profile within the GPFS. With some non-essential for this discussion details, the GPFS with multiple synaptic sites can be considered as a branching array of MOS transistors that have a common drain and individual gates and sources, associated with each synapse. Such consideration can only be made for conceptual purpose, since the proposed structures do not have the source and drain diffusion areas, which are present in MOS transistors. Such structure can easily be fabricated within polysilicon. Therefore, with some simplifications, each synapse can be thought of as a polysilicon FET with source, drain and gate potentials defined by externally applied electric fields. Because the body of the GPFS is DC-isolated from the rest of the electronic circuitry and other distant GPFS', the connections between different devices are achieved exclusively via capacitive coupling.

A potential, applied to each "gate" node, creates an energy barrier between the adjacent "source" and the common "drain" areas (CDA). This "gate" potential is effectively a synaptic weight coefficient that controls the "source" to CDA diffusion current within each synapse of the GPFS. The diffusion currents of all synapses collectively affect the rate of charge accumulation in the GPFS' CDA. The CDA integrates charges from all connected synapses. The total charge represents the weighted sum of all the inputs. Without externally applied electric fields and discounting boundary conditions, the charge within the GPFS is evenly distributed in the CDA. A small strategically selected section of the CDA acts as a gate for the charge sensing Field Effect Transistor or the Sensor Node (SN) located underneath the CDA in the single-crystal substrate. The local charge density in the CDA over the Sensor Node controls the current through the SN. The source drain-current through the SN is integrated in a distant CCD-based (Charge Coupled Device) storage well. The SN acts as a charge-injecting device for the subsequent Charge Pump (CP). A current pulse is triggered every time the amount of charge in the storage node of the CP exceeds a threshold, and as a result the Storage Node is drained. The sequence of these events creates an output train of pulses with frequency and duty cycle being a function of the charge density above the sensor node and control voltage applied to the CP.

The GPFS does not have to be fabricated in a single crystal material and is therefore more tolerant to defects. This structure can be used to create 3D configurations with enhanced interconnectivity. Such structures do not generate excessive heat, and can be fabricated in multiple polysilicon layers; as a result complex neural networks can be made on a single die.

III. Architecture and Operational Details.

The artificial neuron is described further in terms of its two fundamentally different processing domains: 1) GPFS and 2) single crystal substrate. The use of GPFS allows some properties of synapses, dendrites, propagation delays within the "soma" and functionality of the axon hillock to be simulated. Single crystal substrate is used only to produce structures of the neuron that are responsible for a) adjusting sensitivity of the neuron via charge injection and tunneling into and out of the GPFS, b) energy characteristics of the axon, c) shape and nature of presynaptic activity, d) axonal propagation delay and e) speed of "action potential". The following sections describe all these components in greater detail. Although several material combinations could be used to fabricate this device, the discussion is focused on silicon-based materials.

A. Charge Accumulation

A Silicon Neuron, with GPFS as a storage and computational medium, is based on the ability of an electric charge, injected into the floating gate structure through the insulating SiO_2 or Si_3N_4 layer "Fig. 1", to be preserved inside this insulating shell. Application of an external electric field to this structure causes charge diffusion toward the areas with the lowest possible energy, resulting in redistribution of charge density.

By applying appropriate voltages to the respective "Injector" and "Drain" nodes, electrons can be forced in and out of the GPFS. This technique is very well described and has

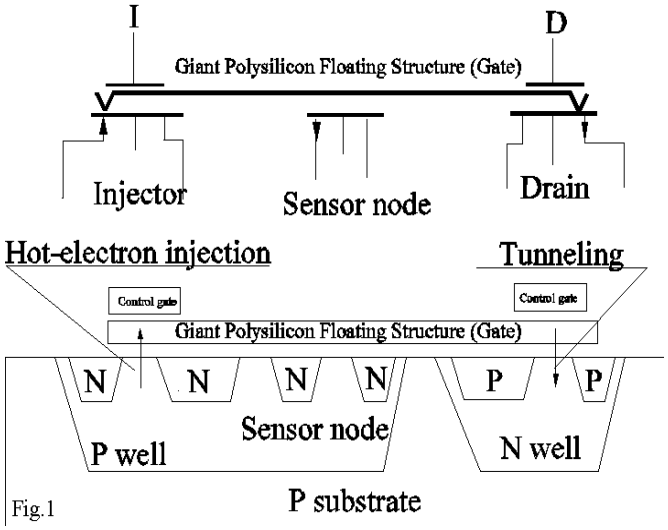

Fig.1

been implemented in designs of Synapse Transistors [9]. A FET (Field Effect Transistor) structure, serving as the charge-sensing node - SN is fabricated below the GPFS. After being injected into the GPFS, and in the absence of external electric fields, charges disperse evenly across the volume of the polysilicon structure. Any change in charge density over the SN also causes the current in the channel of the corresponding FET to change.

B. Giant Polysilicon Floating Gate Structure

With some degree of simplification of Boltzmann statistics, the behavior of charges in the bulk of GPFS under the influence of locally applied electric field can be described in terms of a hydrostatic model. "Fig. 2" presents such a model in a form of a pool that contains a pressure sensor (equivalent to a charge sensor node below the GPFS) built into the bottom of the pool, and several water filled cavities. Multiple pistons are located in the cavities in the bottom; each cavity-piston unit acts as a synapse. As pistons move, water is displaced from cavities resulting in constantly changing level above the pressure sensor. The pressure sensor controls the level of water in the pool and acts similar to the FET Sensor Node in electrostatic case. A precharged GPFS with an associated sensor node and multiple capacitively coupled electrodes act as an electrostatic analog of the hydrostatic model. This structure works as a summing device for applied potentials (piston displacements).

The effect of each synapse on the neuron's overall activity can be expressed by the equation $\mathbf{S_i}=\omega_i\mathbf{A_i}$, where $\boldsymbol{\omega_i}$ is the synaptic weight and $\mathbf{A_i}$, is a presynaptic activation (excitation or inhibition). Presynaptic activation can be thought of as a magnitude of electric field created in the GPFS by voltage, applied to the corresponding control electrode. The field profile induced by several electrodes shape a lateral energy gradient along the GPFS as shown in doted lines on "Fig. 3".

The first component of the synaptic equation is the presynaptic potential $\mathbf{A_i}$. There are several mechanisms

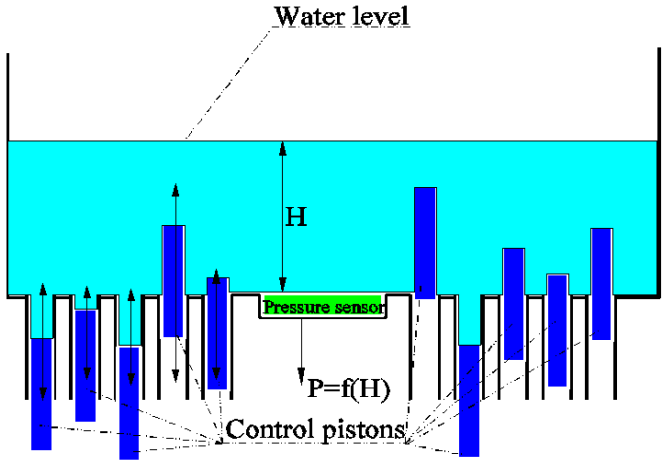

Fig.2

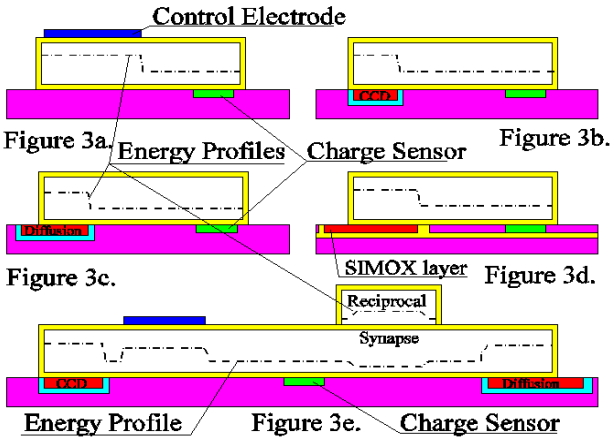

Fig.3

available for simulating the membrane's postsynaptic activity. These mechanisms are based on different configurations and physical structures of the Control Electrodes. All these mechanisms provide some sort of charge biasing (equivalent to the piston's displacement in the hydrostatic model) in the GPFS. Five examples of these mechanisms are: "Fig. 3a" -a capacitively coupled electrode on top of the GPFS, "Fig. 3b" - CCD well filled with electrons and located below the GPFS in the substrate, "Fig. 3c" - a floating diffusion diode underneath of the GPFS, and "Fig. 3d" - a floating gate structure fabricated in SIMOX layer below the GPFS. A variety of other mechanisms are also possible. Several mechanisms can be combined to implement complicated field structure, resulting in even more intricate behavior of the device; In addition to mechanisms listed above, "Fig. 3e" shows a structure utilizing a reciprocal synapse. This type of synaptic interconnection produces bi-directional influence of the local charge density in one GPFS on distribution of charges in another.

The second component of the synaptic equation is the adjustable weight coefficient $\boldsymbol{\omega_i}$. A nonvolatile semiconductor memory element, for example a charge trapping or floating gate device, can be used to implement a long-term memory function for storing synaptic weight coefficients $\boldsymbol{\omega_i}$. "Fig. 4" shows a longitudinal section of such structure that resembles a single

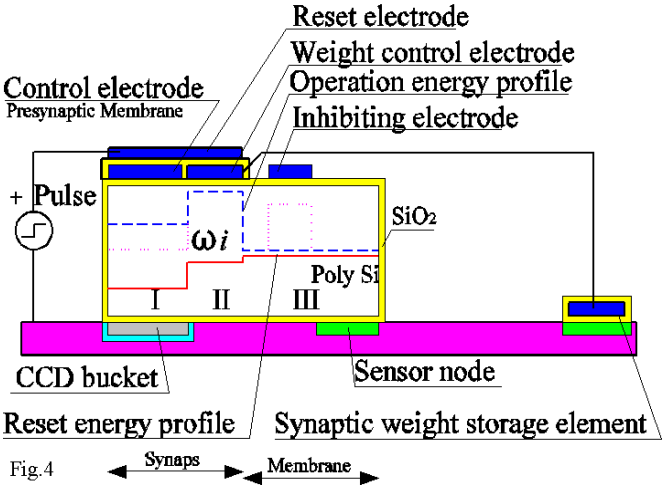

Fig.4

synapse and a portion of the neuron's GPFS. This structure performs low-level signal processing that can simulate behavior of a single biological synapse. The GPFS is encapsulated in SiO_2 shell and consists of a number of synapses and the common summing node. Each synapse (excitatory or inhibitory) consists of two areas: 1) Area "I" where the presynaptic potential effects the postsynaptic energy profile and which is equivalent to $\mathbf{A_i}$ in the synaptic equation, 2) Area "II" where presynaptic weight adjustment occurs and which performs the multiplication by $\boldsymbol{\omega_i}$, and 3) Common Drain Area "III" where the charge summation is taking place and which mimics the behavior of a biological neuron's membrane. In other words Area "I" is equivalent to an excitatory synaptic junction where neuro-transmition occurs, Area "II" defines the rate of "synaptic metabolism", while Area "III" is the summing node, which integrates efforts of all synapses to depolarize the membrane and trigger a neuron's firing.

As the result of prolonged operation of the synapse, Area "I" becomes depleted and charge diffusion from "I" into "III" degenerates. This is equivalent to a neuron's hyperpolarization. When this condition occurs, high positive voltage is applied to reset the synapse to its operational condition. Another possible mechanism is associated with a constant, but very small tunneling current, flowing into each synapse; this current is being integrated in the synaptic storage node, and the produced charge is rapidly drained-out during neuron's firing. In which case an extra time is required to restore the minimum amount of charge. "Fig. 4" shows corresponding structure designed to simulate functionality of an excitatory synapse. An inhibitory synaptic function may be added as needed to an individual excitatory synapse or to a branch of excitatory synapses - dendrite.

C. Somatic Mechanism and Charge Pump

The somatic mechanism includes a Charge Sensor, VCO (Controlled Oscillator) equipped with a threshold control mechanism or Charge Pump "Fig. 5", and a negative feedback somatic loop shown on "Fig. 6". This structure is designed to generate an appropriate oscillatory response upon excitation of the dendritic tree. It also prevents the Neural Circuit from being overexcited through the introduction of local negative feedback. Additionally, threshold control simplifies the process of network adjustment to different levels of background activity.

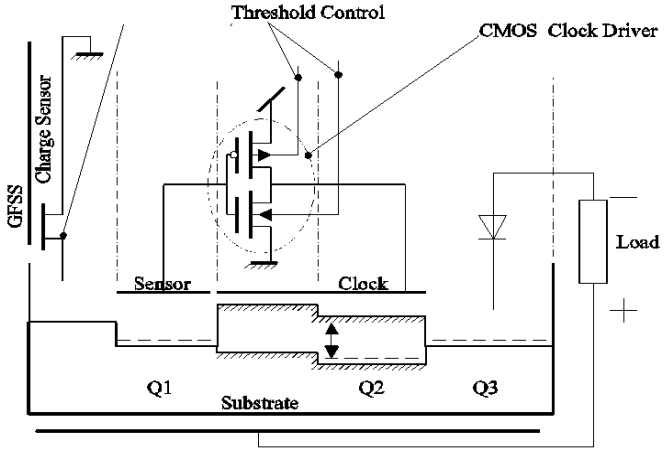

Fig.5

"Fig. 5" shows the structure of a simple device that simulates firing activity of a neuron and performs functions of a VCO - the Charge Pump (CP). This structure is essentially a small CCD shift register (four energy wells) along with some associated control circuitry. A current, produced by the charge sensor (injector), introduces charge into the CCD well Q1, this well has a specified charge-holding capacity defined by its size and doping concentration. Voltage on the sensor electrode or the gate of the CMOS driver depends on the amount of charge in this well. When the sensor voltage exceeds the threshold of the driver, the output voltage of the CMOS changes from low high and forces the charge from Q1 to Q2.

Hysteresis-like behavior of the CMOS driver is important to prevent high frequency oscillation of the output of the device around the point of equilibrium. Hysteresis is achieved by additionally biasing the bulks of NMOS and PMOS components of the driver. When high potential is applied, the lateral field under the clock electrode forces charges to move from Q1 to Q2. As the result, voltage on the sensor electrode drops and the CMOS returns to the state with low voltage output, which elevates the energy profile in Q2 and prevents further emptying of the Q1. At the same time, because of the graded doping profile in Q2 bucket, the charge from Q2 is forced into Q3, where it is drained through the diode/resistor circuit, creating a voltage drop across the resistor. When the voltage on the clock electrode drops below a certain level, charge from Q2 moves into Q3, and Q1 begins the next cycle of charge integration (accumulation). The diode connected to the Q3 node allows the charge to be drained into the load. This process repeats continuously for as long as the injector injects electrons into Q1. The duration of pulses depend on the channel current through the injector. Besides providing a hysteresis-like behavior of the pump, the threshold control mechanism allows additional degree of freedom for adjustment of the neuron's sensitivity.

When the neuron is in a state of low background activity (subthreshold state), the charge pump fires with relatively low frequency and high pulse duration. Any firing at all is due to the always-present parasitic current through the channel of the Charge Sensor FET. Voltages applied to threshold control terminals of the Clock Driver set the point of the background activity of a neuron at specified level. The CP starts firing with lower duration (higher frequency) when the output of the charge sensor FET exceeds the threshold value. The frequency of spikes gradually increases (duration decreases) in response to higher current through the Charge Sensor FET. In some cases, the length of the CCD-based charge pump can be increased to introduce an additional signal delay. In other cases, the firing rate of one neuron may control the propagation delay in the other neuron by clocking its charge pump (Axon).

D. Overall Operation of the Silicon Neuron.

The GPFS with embedded synapses, Charge Sensor Node, and associated charge pump form a structure that simulates some functions of a biological neuron. In cases where there is no need for propagation delay, a simple conductor is used to convey firing of one neuron to other neurons; if a delay is required - a CCD-based axon is introduced into the signal path. Use of CCD opens possibilities for both spatial and temporal signal processing; the later is due to controlled propagation delay - when one neuron controls the rate of firing activity of a distant neuron by setting up a CCD clock frequency.

CP generates pulses that are applied to inhibitory and/or excitatory synapses of distant neurons "Fig. 6", and control charge buildups in their corresponding GPFS'. Sometimes it may be necessary to introduce a negative-feedback local somatic loop. Such a loop employs direct output from the CP or from the CMOS sensor to trigger a blocking mechanism. When the neuron becomes overexcited, an input voltage on the CMOS sensor increases, which causes increased firing activity of the CP. A dedicated inhibitory synapse or a group of activity blocking inhibitory synapses remove some electrons from the floating structure, and by this means keep the neuron away from the overexcited state.

Fig. 6" shows five synaptic organizations integrated into a single Silicon Neuron: 1) A synapse with overlaying Control Electrode, similar to "Fig. 3a," 2) CCD-base charge (stimuli) transporting structure for presynaptic temporal activation, similar to "Fig. 3b," 3) A synapse with an underlying diffusion area in the substrate Control Electrode, similar to "Fig. 3c," 4) Combined synapse with both overlaying and underlying Control Electrodes, similar to "Fig. 4," 5) a reciprocal synapse with several GPFS's from different neurons overlaying each other and possibly presynaptic membranes of Type 1, 2, 3, and 4 over the synaptic site to form a complex interrelated arrangement, similar to "Fig. 3e." In addition to different types of synaptic organizations "Fig. 6" shows the structure of a Dendritic Trunk with Dendritic Inhibition, Axon Hillock Inhibition, somatic

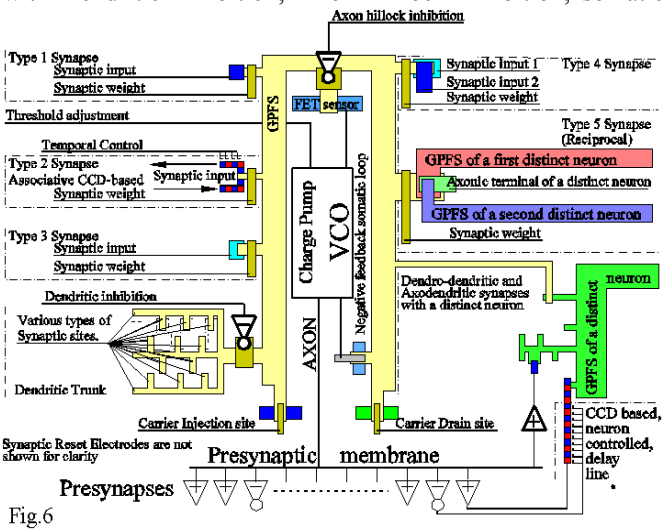

Fig.6

mechanism, and an Axonal propagation delay mechanism (a CCD-based inter-neuron controlled delay). As a separate insert the simplified graphical representation of several Dendro-Dendritic and Axo-Dendritic connections is also shown.

E. Axonal Propagation Delay Mechanism.

The main purpose of the axon is to deliver the firing activity of a neuron to a presynaptic membrane of a distant neuron. In some cases, like in processing of spatial visual information in the retina, LGN (Lateral Geniculate Nucleus), olfactory and taste nuclei, etc. temporal processing is not required; in other cases, like speech processing and synthesis, motor actions, and any other processing requiring time-coordinated efforts from many neurons, such a capability becomes highly desirable. It is becoming essential to control the signal propagation delay in neural networks for temporal processing. In the proposed structure of the Silicon Neuron the delay can be introduced via implementing a CCD-based signal-transporting scheme in the structure that simulates the behavior of an axon.

There are at least two mechanisms available to control the propagation delay of the firing activity of one neuron to another: 1) self-induced delay, and 2) externally induced delay. Since the Charge Pump is essentially a CCD structure that produces a PWM (Pulse Width Modulated) sequence, the delay is automatically introduced as a function of the neuron's firing frequency and the length of the CCD register, which is equivalent to the number of steps from the beginning of the axon (cell body) to its end - synaptic cleft with the target. As a result, extending the Charge Pump structure and adjusting the threshold of the CMOS clock driver - "Fig. 7", allows manipulation of an internally induced delay. Self-induced delay is also a function of the postsynaptic activity of the Silicon Neuron.

If needed, an externally induced delay is implemented by utilizing a structure similar to the one shown on "Fig. 7", but instead of clocking the CCD from its own Charge Pump, the propagation of charge packets along the axon of one neuron is controlled by the firing activity of another neuron. Under these conditions the signal propagation speed can be adjusted

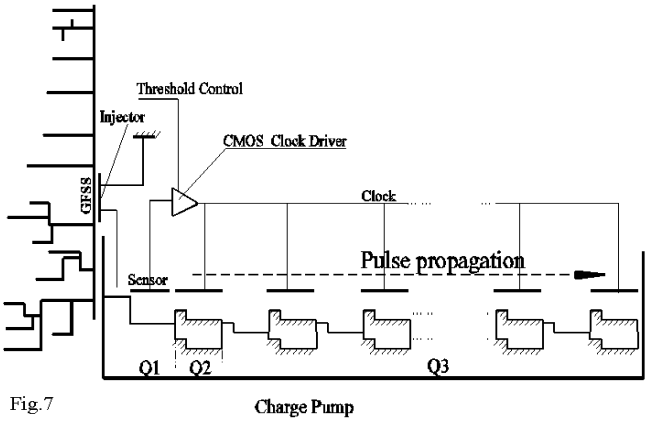

Fig.7

according to the needs of a particular network.

The delayed firing of a neuron can also be utilized to construct ANN's exhibiting behaviors similar to Hebbian oscillations. Such oscillations are usually associated with Reverberating Cell Assemblies and require delayed feedback loops; the example is presented on "Fig. 8". Neurons 1, 2, and 3 are the components of the reverberating Cell Assembly, similar to Hebb's discription; the assembly receives an input stimulus from a "Mossy Fiber" arriving from a distant location in the neural network (cerebral cortex). Firing activity of Neuron 1 is controlled by an externally induced delay, which is produced by the Neuron 2. Firing of Neuron 2 is defined by its presynaptic activation and is not delayed, while firing activity of Neuron 3 is defined by its presynaptic activation and is delayed in accordance to its activation (self-induced propagation).

Multiple feedback control loops may exist in such structures that allow desirable signal conversion and conditioning. For example, a short stimulation of a "Mossy Fiber" may trigger prolonged oscillations of the assembly by initially exciting all three neurons of the assembly, followed by delayed firing of Neuron 1 that reinitiates the process. A delay introduced into the positive feedback loop, that includes Neurons 2 and 3, maintains the oscillations for a duration that can be defined by the characteristics of these neurons.

IV. Simple Network Implementation.

Behavior similar to the one described above can be used to simulate the activity of some of brain's auditory functions. The auditory cortex processes information, that was converted by cochlea into the frequency domain [10], so that fluent speech can be filtered from noise and decomposed into simple components, similar to phonemes. The representation of each phoneme may exist in the brain's Broca's area as a hardwired Reverberating Cell Assembly. When exposed to a speech sequence, a unique, phoneme-specific temporal assembly becomes excited only if the corresponding phoneme is present in the auditory stimuli. At the same time, a similar structure may exist in Wernicke's area, and when activated by a simple short stimulus, it produces a synchronized pattern of firing. This pattern stimulates various facial and glottal muscles to produce the sound of the desirable "hardwired" phoneme. Similar structures can also be expected in different locations in the brain responsible for memory, reflexes, motor control, etc.

V. Conclusions.

This paper presents an approach to building semiconductor computational elements that closely resemble topologies and behavior of biological neurons. It uses standard semiconductor fabrication technology in an unorthodox manner to build neuromorphic computational components that are not based on traditional models of MOS and BJT transistors. Controlled electrostatic fields applied to a precharged poorly conducting medium force redistribution of charges in the bulk of it and

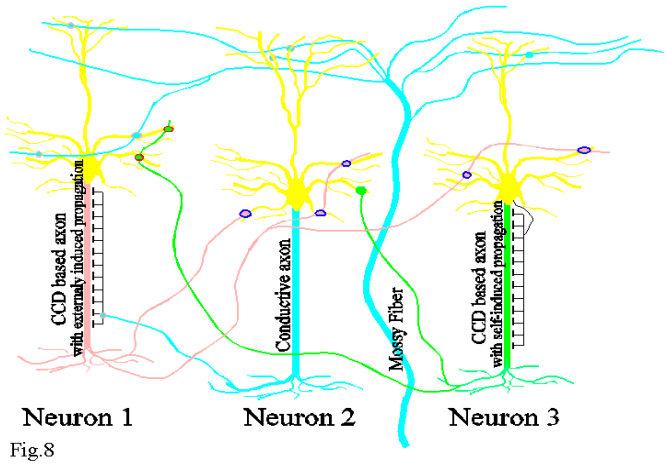

Fig.8

make complex computations possible. Physics of the proposed device, its architecture, and the material used allow stacking of multiple layers of processing elements and 3D interconnects on a single silicon substrate. High density interconnects, no-current operation and neuron-oriented computations allow for a higher level of system integration and thus, building more complex nets on a silicon die. However, many intricacies of the device operation are not yet fully understood and require further investigation via simulations and experimental fabrications.

VI. References

[1] Hebb, D. O. (1949), The organization of behavior , New York: John Willey

[2] R. Douglas and K. Martin, (1998), "The Synaptic Organization of the Brain", Fourth Edition, Oxford University Press

[3] William H Calvin, (1996,) "The Cerebral Code," MIT Press.

[4] Christof Koch, (1999), "Boiphysics of Computation", Oxford University Press

[5] Carver Mead, (1989), "Analog VLSI and Neural Systems," Addison-Wesley Publishing

[6] Yoshihiko Horio, (1992), "Analog Memories for VLSI Neurocomputing" in Artificial Neural Networks, IEEE Press, pp. 344-363.

[7] Alan F. Murray, (1989), "Pulse Arithmetic in VLSI Neural Networks," IEEE Micro Mag., Dec. 1989, pp. 64-67

[8] Michael C. Mozer, (1993), "Neural Net Architectures for Temporal Sequence Processing," Institute of Cognitive Sciences, Univ of Colorado

[9] P. Hasler, C. Diorio, B. A. Minch, and C. Mead, "Single transistor learning synapses with long term storage," IEEE Intl. Symp. on Circuits and Systems, vol. 3, pp. 1660–1663, 1995

[10] V.A. Gorelik, "Hypothetical Mechanism of Auditory Processing for Extraction of Directional cues and Integration with Oculomotor Function," unpublished