

The vWISP Neural Processor.

White Paper: Introduction to a Neuro-computing device

Updated: 21th August 2008

vWISP Pty. Ltd.

Peter AJ van der Made

CTO

Key Words:

Artificial Neural Network, Artificial Brain, Bayesian Inference, Temporal Associative Memory, Event driven systems, Stochastic learning systems, Synaptic Time Dependent Plasticity STDP, Markov chains.

Introduction.

It is difficult not to be impressed by the capabilities of the human brain; we can instantly and effortlessly recognize people, places and circumstances. Most of us can also figure things out from imprecise or confusing information, make a decision based on reasoning, act on that decision and learn from the experience. A brain can, and does, lose thousands of neurons and continue to function as if nothing has happened. In contrast, all these are things a computer cannot do, or at best do poorly with the aid of complex and task-specific programs. In addition, a computer stops working if a single one of its millions of logic gates fails.

People have been studying the brain for over 150 years. There is no shortage of documentation on the brain. Several mountains of paper document all aspects of neuron-physiology, pathology, electro-physiology and the function of specific brain regions in amazing detail. Some of those papers appear to contradict each other, or approach specific neural functions from a different and confusing point of view. Many hypotheses get lost in accumulated complexity offering too much detail, mathematical theories and algorithms, but offer few or no concrete conclusions. The simple fact is that the brain does not execute any code nor does it run algorithms – The opinion expressed here, and the basis for the development of the vWISP neural device, is that everything the brain does needs to be explainable from neuron physiology and the synaptic connections between neurons. I am convinced that the connectivity of neurons and synapses holds the key to intelligence. Those who do not believe that we can form a unifying brain function theory based on the brain's smallest functional components refute to comprehend neuron

physiology. Mathematical models, while instructive, will never be able to be processed in real time on any computer system in sufficient numbers, at least not one that fits into a moderate size living room.

Therefore, the approach that is taken here is to construct the complex functions of the brain from biologically realistic, but digital neurons and synapse functions, hence build an artificial brain that functions in the same manner as a biological brain. We must be able to accept that as we emulate brain function, eventually we will face the problem that we can not control what this brain learns. Independent thought and learning are build into the hardware of our brains. In small arrays of a few million neurons this is not going to be a problem.

Due to the availability of better imaging and research tools, new discoveries in the neurosciences have opened up avenues to simulate accurately the function of neurons and synapses. The algorithm is thought to be similar to Bayesian Inference, a mathematical method to calculate probability based on previous proof. Several software projects have been initiated in an attempt to build intelligent systems based on Bayesian Inference. However, software is limited to the capabilities of computers which fall short of simulating the complete brain. Bayesian inference alone can not account for neuron function without a time dimension, feedback, association and the constant learning or reinforcement of activation patterns. Interaction between neurons may be explained using swarm intelligence algorithms. Bayesian Inference is part of the puzzle, not the whole solution.

In the model that is presented here the brain is perceived as a massive event-driven system, whereby sensory neurons fire events which are processed in parallel in a network of LIF (Leaky Integrate and Fire) neurons with dynamic synapses. Each layer of neurons evaluates these events as temporal probabilities, reducing data intensity and increasing sophistication as data travels up the hierarchy. Each output is a deduction of combined temporal and spatial input events. These dynamic artificial neurons (DAN) form a tree structure of massive parallel, event driven, STDP-learning Temporal/Predictive nuclei with feedback and association. The neurons interact by simple rules like a swarm of nuclei. We focus on rules of interaction rather than connections.

Any brain algorithm that ignores neuron function may produce short term benefits, but it unlikely to scale to higher brain functions and is in danger of eventually collapsing in complexity as more advanced features are added on.

Previous Attempts to simulate brain function.

Due to its complexity the brain has often been approached from the perspective of a (over)simplified model and compromises have been made to get the system to work with the tools that were available at the time..

Previous artificial neural networks; Perceptron¹, Hopfield² and Kohonen self-organizing nets are simplifications of the findings of early brain researchers. For the larger part such networks have been implemented in computer software and are poor simulations of brain function, missing significant temporal dynamic properties and feedback connectivity, which are important factors in neural information processing. Most of these earlier neural nets assigned static values to synaptic inputs, which are then added in a soma whenever an input is active. The static value represents the weight of the synapse. Each static 'neuron' connects to every other static neuron. Such limited implementations are doing useful work in pattern recognition and forecasting, but perform poorly or not at all when scaled to higher functions. AI research has long been impaired by attaching too much importance to rigid models which have no contemporary biological basis, and by teaching students that computers are capable of emulating the brain.

In later developments the synapses are updated and are referred to as 'Dynamic Synapses'. The Berger-Liaw neuron model uses a mix of digital and analog components to create an accurate model implemented in VLSI³. The model consists of a signal processor to simulate neural function and analog 'processing junctions'. One projected application of this technology is to produce implantable electronic devices, such as an artificial Hippocampus for sufferers of brain damage⁴. The Berger-Liaw design has attracted significant publication interest when it was able to recognize words embedded in a high degree of noise – yielding better results than could be attained by human hearing, but with a vocabulary of only a few words.

Some models ignore the neural function completely and use Bayesian Inference to simulate the higher-level function of a group of neurons, whereby Feedback is a function of Evidence. The Hypothesis is the probability that the input pattern converges with the pattern stored in synapses. However neurons are not pattern recognition engines, but association engines. They associate the input pattern by means of the strength stored in synapses, not by pattern matching.

Problem definition.

In the brain, many millions of events are handled immediately and simultaneously. There is no separate memory block that needs to be addressed to retrieve or store information.

¹ Frank Rosenblatt, Cornell University, 1960 - "Principles of neurodynamics: Perceptrons and the theory of brain mechanisms" (Spartan Books, 1962)

² John Joseph Hopfield, Hopfield Network 1982 "Neural networks and physical systems with emergent collective computational abilities"

³ IEEE Computer Society, Theodore Berger and Jim Liaw et al, USC 2000 "Compact VLSI Neural Network Circuit with High-Capacity Dynamic Synapses"

⁴ Theodore Berger et al, USC, 2001, "Brain-Implantable Biomimetic Electronics as the Next Era in Neural Prosthetics"

Information is stored within the synapses of event-handling neurons themselves. No 'central processing unit' exists. There are no bottlenecks in this processing array.

In contrast, computers are serial devices which fetch instructions from a serial program list and process information stored in random access memory sequentially. This sequential process loop is to fetch an instruction, decode that instruction, retrieve the address of data, and to fetch that data from memory or an input peripheral. The processed data is written back to memory or an output peripheral in subsequent instructions. Program branches and jumps perform 'skips' in a serial list of instructions. Parallelism is simulated in these machines in several ways. One is to execute several 'treads' from a treading pool, whereby the central processor switches between the treads. Each tread is a semi-independent section of a program instruction list. It appears as if treads run in parallel, but really each tread is executed sequentially and the switching occurs very fast. In addition to the treading model, most computer systems support hardware and software interrupts. Interrupts disrupt serial program execution and pass control to an interrupt service routine, as a means to process time-critical events. Tread-switching is often performed through a dedicated interrupt routine.

Due to the inherently time-sequential method of execution there is a limit to the number of treads and interrupts that can be executed and appear to run in real-time. In a hypothetical neural simulation program each neuron could be assigned to a tread. Ideally, each synaptic input triggers an interrupt, and interrupts are shared. The interrupt service routine would need to determine which input triggered the interrupt, precisely time the input pulse and the resulting increase in membrane potential, and determine the increase or decrease in synaptic strength as a function of output pulse to input pulse timing. Within this tread the soma activation history needs to be stored to calculate the threshold offset. The membrane potential is stored and decreased according to a logarithmic function over elapsed time. The synaptic strength is stored and constantly updated. In a compiled and optimized C program this function took an average of 700 μ S on a 3 GHz dual-core Pentium computer. Three parameters need to be stored for each neuron and at least one for each synapse.⁵

All active neuron treads need to be processed, even when the tread is not part of a critical input-to-output path. At a reaction speed of 200 mS., a maximum of 285 neurons can be processed in this manner. To simulate the intelligence of a cockroach a system consisting of thousands of parallel processors would be required.

These disparities between the brain and computers thus form a bottleneck that becomes increasingly problematic when large numbers of neurons need to be simulated.

Dr. Eugene Izhikevich (a scientist at the Neural Sciences Institute, San Diego CA) has performed a one second simulation of a pulsed neural model of a brain with 100 billion neurons on a Beowulf (network) cluster consisting of 27 Intel Pentium-4 computers

⁵ Eugene Izhikevich 2003 IEEE transactions on Neural Networks "Simple Model of Spiking Neurons"

running at 3 Ghz.⁶ . This one second simulation took 50 days of computer time to execute. His simulation software was compiled C code. In other words, this large cluster of computing power would **need to speed up by a factor 4.3 million to simulate brain activity in real time** (50*24*3600) seconds * 27 computers). A single processor would need to speed up by a factor 116,640,000 times. The simulation took an average of 1.16mS per neuron (time x computers / number of neurons= $1.16 \cdot 10^8 / 10^{11}$. Eugene has a table of processors vs. processing clock speeds on his web site: see (<http://vesicle.nsi.edu/users/izhikevich/interest/why.htm>).

Even when using many fast processors (over 58 million processors, running at 6 GHz) it is clearly not realistic to simulate the entire human brain. Why use software to make a system perform a task to which it is not suited? A new hardware architecture is needed to process millions of concurrent temporal neural events.

Therefore, our approach has been to build a parallel array consisting of digital circuitry that simulates the neural function with biologically realistic behaviour. The array does not use software or a traditional central processing unit. The premise of the vWISP architecture is that, if a digital system is built according to conventions defined in a working biological system, then the complete digital system will perform the same function as that biological system.

Each new version of this digital system has been a refinement on the previous version, and closer to the function of the biological model. By examining the behaviour of the model, we have gained a better understanding of the biological model, which is incorporated in the latest technology. We expect that this process of refinement will continue into the future.

The NeoCortex-1 processor has been superseded by the NeoCortex-2, which incorporates SDTP learning and a variable soma threshold in addition to the leaky integrator, feedback and dynamic synapses found in early pre-release versions and the NC1 chip.

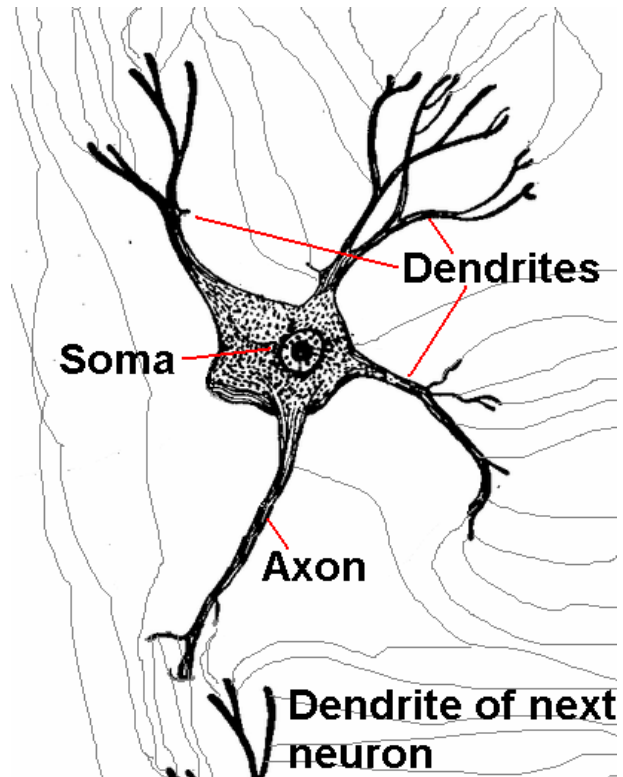
Rather than determining the exact connections that have evolved over time in a fully mature brain, the array is a swarm of neurons and the interactions between neurons is defined. Each neuron interacts with many other neurons according to simple rules and thus a complex system evolves over time. Each neuron is a Bayesian Temporal association engine with many inputs that constantly update or reinforce their activation strengths. Not every neuron interacts with every other neuron, but like in a swarm of bees with a specific selection of them and according to simple rules. This would make a chaotic system were it not for a series of delays that organize and synchronize the array. Asymmetric feedback reinforces or inhibits learned patterns. Prediction and association are directly related to the combination of learned Bayesian inference, swarm intelligence and feedback.

⁶ Eugene Izhikevich 2007 Neuroscience Institute, John Hopkins University, "Large-Scale Model of Mammalian Thalamocortical Systems"

The Biological Brain.

A textbook picture of the brain shows a pale pinkish, wrinkled outer layer. This outer layer is the neocortex, consisting of six layers which are divided into sub-layers. Most inputs come into layer 4, which consists of stellar neurons and projects to pyramidal neurons in layer 2. Mapping connections between layers, feedback and connections to sensory neurons and motor neurons rapidly starts to look like a large tub of spaghetti. The structure of the neocortex is complex, but looks the same everywhere. The vision and auditory cortex have the same structure, and so do the motor cortex and the prefrontal cortex. The prefrontal cortex is where we do all our thinking. Because of this homogeneous structure, Vernon Mountcastle made the assumption in 1978 that the pattern recognition and prediction algorithm in the brain is the same everywhere, an assumption that is now widely accepted and supported by evidence.⁷

Each neuron is a dynamic temporal pattern recognition engine. Input to a neuron is through synapses. Synapses cover the dendrites, but may form anywhere else on the neuron including the soma and the axon. The closer to the soma the synapse occurs, the greater its effectiveness. Synapses are connected to the axon output of other neurons and sensory neurons. Synapses receive pulses, and each pulse results in an increase in Post Synaptic Potential (PSP), which are integrated in the cell membrane. The membrane potential is constantly declining toward the rest potential. The neuron 'fires' e.g. produces an action potential when the membrane reaches a threshold potential. Feedback strengthens a synaptic connection if the pulse occurred before the output pulse and weakens synapses that were pulsed after the output pulse occurred. In this manner the synapses learn to recognize, and later reinforce a repeating pattern. New synapses form constantly. It is assumed that new synapses form as a function of permanent memory, while the strength of synapses is increased or decreased as a function of temporary memory.



The central nervous system exists out of an estimated 100 billion (10^{11}) neurons interconnected through an average of 7000 synapses each, and with 1 trillion (10^{12}) glial

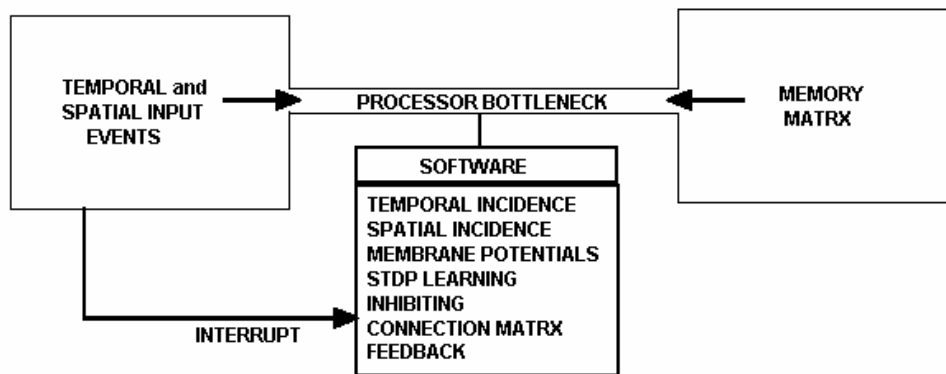
⁷ Vernon Mountcastle, 1978 "The Mindful Brain", MIT Press

support cells. There are an estimated 100 trillion (10^{14}) synapses in the brain. A biological neuron is a slow device. Transfer of an input to an output event can take between 5 to 40 milli-seconds. Silicon NAND gates are about 5 million times faster, responding to an input in 2 nano-seconds. Yet, due to their massive parallel architecture, brains win on most fronts.

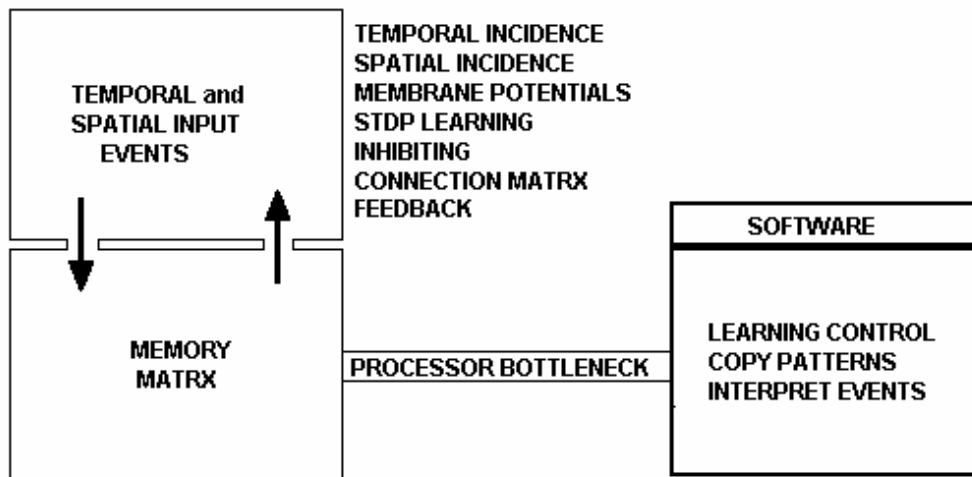
The vWISP hardware solution.

The vWISP neuron model simulates the properties of a neuron that are currently known from experimentation with biological neurons. It accomplishes this by using simple logic gates, registers and counters in a custom design. The neuron consists out of the equivalent of 3500 logic gates. A matrix of these neurons forms a hierarchical processing array. The synapses are dynamic, e.g. they learn new values, or reinforce a learned value over time. Synapses are strengthened when they contributed to the output pulse and weakened when they occur after the output pulse simulating a process known as Synaptic Time Dependent Plasticity.

This array processing method is highly flexible, but is not programmed in contrasts with traditional 'serial' computer designs. Serial execution units need to fetch instructions from memory, retrieve and store data in memory and process interrupts sequentially. In this way the processor forms a bottleneck through which all events have to pass:



We can avoid the processor bottleneck if we input straight into memory. This idea of course has been around for a long time, called Direct Memory Access or DMA. However, this is a very different technique whereby the information is processed immediately and within the same device. The array looks like memory to an external microprocessor. The microprocessor can read the state of any synapse and neuron at any time.



In the vWISP NeoCortex processor, temporal and spatial events are processed directly without passing to memory. Incoming events are integrated in a leaky integrator and synapses are updated depending on input to output pulse timing (STDP). An output pulse or multiple pulses are produced if the membrane voltage reaches the threshold level and the threshold level is updated. The state of the synapses, the activation patterns and the threshold voltages can be read back by an external micro-processor. The external microprocessor does not need to perform any neural processing. Each input pulse is an event and is handled immediately, no matter how many pulses occur simultaneously or within a short time. Not every pulse is stored or results in an output. The prime reasons for the microprocessor interface are threefold;

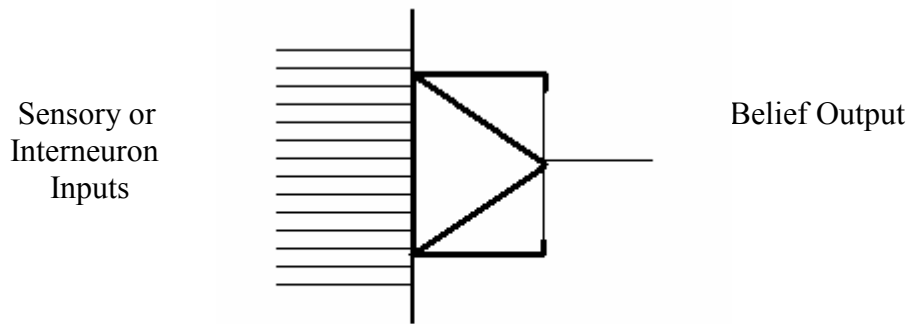
- To allow a standard microcomputer to monitor and interpret the results stored in the neural computing array.
- To allow a microcomputer to clone the results stored in the array. To make a copy of the state of all neurons and synapses once a task has been learned so that this knowledge can be copied to multiple arrays, instantly training those devices.
- To allow a microcomputer to seed the matrix with previously determined values to train the array, or to initialize the array.

Conclusion.

The vWISP neuron is a new processing element that simulates the analog processing functions of a biological neural cell, but is composed of binary gates. The device offers realistic neural processing performance. The device is intended to be integrated in a large matrix and can be produced using current available technologies, in the order of tens of thousands of processing elements per device. The matrix is not a static programmed

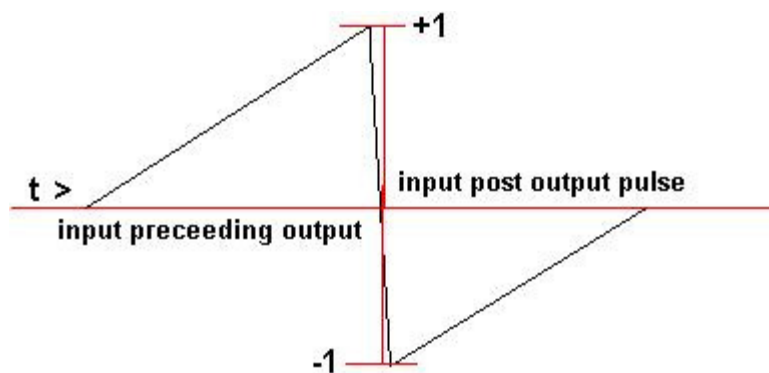
device but a dynamic, constantly learning and self-optimizing system that requires almost no user intervention. A small micro-controller can be used to read the matrix, or to seed the matrix with random or replicated values, but the microprocessor does not perform any neural processing. Once a device is fully trained the microprocessor interface can be used to replicate the learned image from one device to the next.

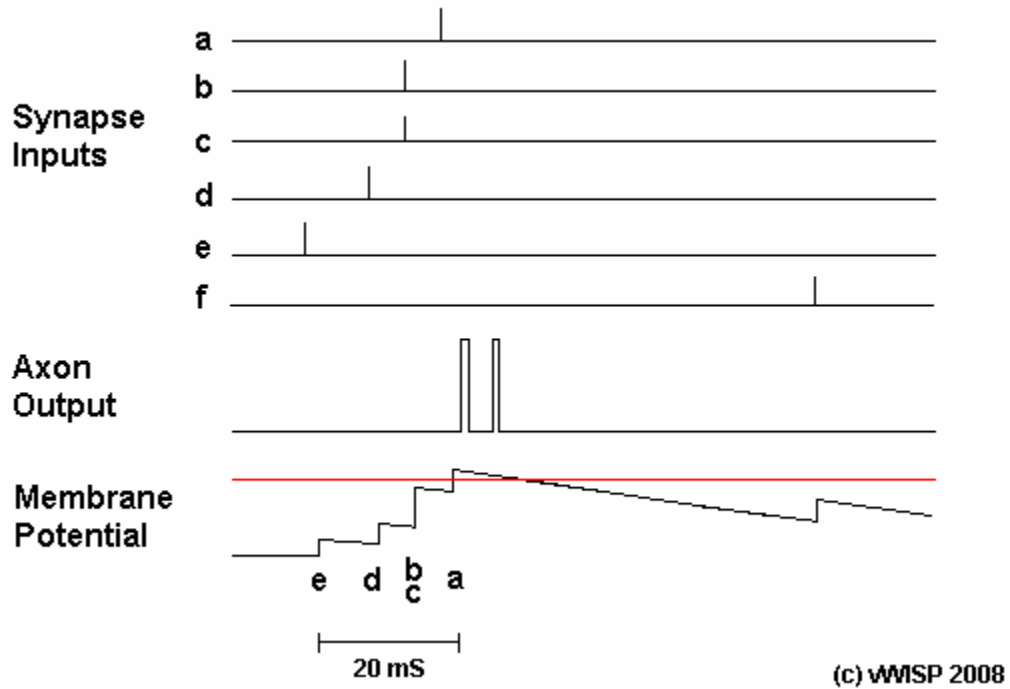
The device does not slow down when large numbers are placed on a single chip. A device employing 15,000 neurons would perform at ~60 times the speed of a 3GHz desktop computer that is executing a neural model.



Dynamic Neuron Logic Symbol

When an input event occurs, the membrane potential is increased by a value that depends on the strength of the synapse. The membrane potential is constantly decreasing and returns to the rest position over time. However, the membrane potential is increased again from this higher level if another pulse occurs before the rest position is reached, resulting in a higher membrane potential than in the first pulse. The time-out value of the membrane is variable, depending on the strength of the synapse. From the maximum level the time-out period is 17.9 mS when the device is clocked at 14.3 KHz or 25.6 μ S if the device is clocked at 10 MHz. The membrane threshold value is eventually attained and the neuron produces an Action Potential output pulse. The strength of all synapses that contributed to the output event is increased by a value determined by the input to output pulse timing. The closer the input pulse is to the output pulse, the greater the increase (max =1 in 16 linear steps). Similarly, the strength of synapses that did not contribute to the Action Potential, which input occurred after the Action Potential, is decreased on a linear scale.





In the diagram above five synaptic input pulses are shown that result in increases in membrane potential. The membrane potential is expressed as a value in an integrator register. Pulses (e), (d) and (a) are temporal integrated and (b) and (c) are spatial integrated. Pulse (a) brings the membrane potential above the current threshold level (red line) and consequentially two fast axon action potentials are output. The strength of synapse (e) is increased by the least amount and synapse (a) is increased by the largest amount. The strength of synapse (f) is reduced slightly, the pulse occurred after the output event and its input did not contribute to the output event.

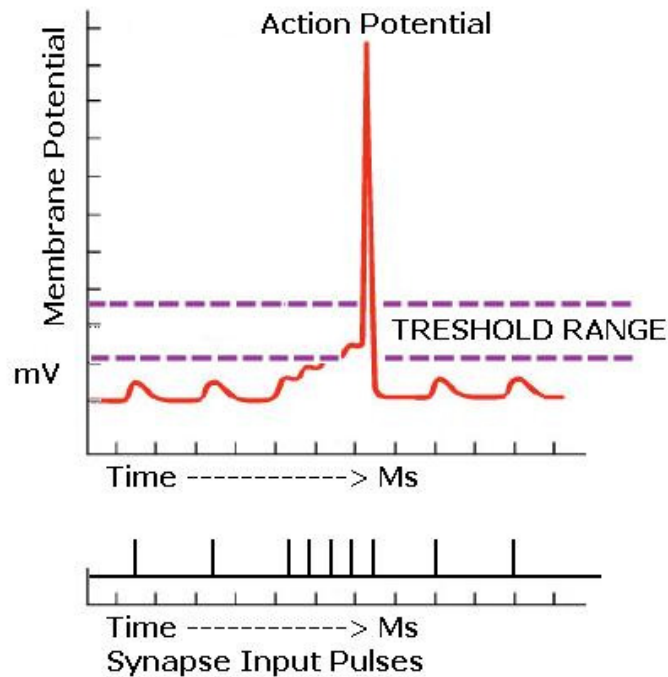


FIGURE 4a Biological Neuron Response

The same behaviour can be observed in biological synapses. The Membrane Potential of a biological neuron is seen to increase with each synaptic input until the membrane threshold potential is reached. At that point the neuron outputs an action potential – a sudden increase in the axon potential that immediately resets. Observe that the area under the threshold level corresponds closely with the “Membrane Potential” graph in the previous image.

The vWISP array processor, like neurons in the mammalian brain, is organized in functional groups, columns and modules. The device is hierarchical; a group fires in a learned sequence and represents a memory and/or an action. A column represents a functional memory segment. Each column triggers the next column before the next stimulus arrives. To understand this mechanism, consider the prediction that takes place while we listen to music; the listener 'knows' what the next tone should be before it is played, even when the listener knows nothing about music. If the wrong note is played, we know. Our brains predict the next note that our eardrums are going to detect by pre-fetching the content of the next column. Modules exist out of groups, which exist out of columns.

At the next level, modules are the partitions that neuro-scientists have defined, such as LGN, V1, V2 V3 and MT in the visual cortex that in combination perform a distinctive

function. These areas were largely defined by examining function loss in persons that suffered severe brain damage and by animal experiments.

Each neuron is a member of multiple groups resulting in a nearly unlimited capacity. In practice, neurons are organized in functional modules and groups that are in close proximity (within several centimeters) which limits somewhat the number of possible combinations. Synapses hold the key to learning. Synapses are dynamic in nature and are constantly updated. The hippocampus appears to perform an important function in learning. Patients that suffered severe brain damage in the hippocampal area were unable to learn new tasks. Rather than concluding that the hippocampus is the 'learning center' of the brain that may indicate that the hippocampus is a control center that allows neurons in other parts of the brain to acquire new data. Thus, the hippocampus appears to have a function in synaptic plasticity.

1. **Learning in the brain is not a simple matter of increasing static weights.** Memories are retained when they are referenced. This appears to be supported by the STDP implemented in the vWISP DAN. Also known as Hebbian Learning, after Donald O. Hebb⁸, it does not explain organization of the brain. The brain appears to contain a self-organizing structure. It is known that traumatic memories are enforced by the effect of adrenalin. Other neuro-modifiers may be responsible for organizing our brains during sleep periods. The 'brain algorithm' is likely a combination of STDP, Kohonen self-organizing maps, Swarm Theory and Bayesian Inference.
2. **Synapses are more than simple connections.** Each synapse is a temporal 'leaky' integrator. In other words, this means that incoming pulses are integrated over time, whereby the integrated value declines in-between pulses. The neuron is triggered when the combined integrated value reaches a trigger value. The trigger value is not constant but depends on the firing history of the neuron, settling at -55 mV over time. Both these mechanisms are implemented in the vWISP Dynamic Neuron. Most neurons also have an upper limit, whereby the neuron trigger is disabled if the pulse frequency is high for a period of 30 minutes to several hours, causing a Ca⁺ deficiency in synaptic vesicles. If effect, a synapse is a filter with memory, with at least two time constants; the preferred interval period between pulses and the recovery time after a trigger event. Both these time constants appear to have an important organizing function in stability in the brain. Refer to work by Walter J. Freeman at UC, Berkeley.⁹
3. **Feedback is asymmetrical,** the output of a neuron can be seen as a Bayesian Inference of its input events, representing that a temporal input pattern, or an approximate temporal input pattern has a certain meaning. This is a 'belief' derived from 'Causes' in Bayesian terms. The meaning of an output event is diverse; depending on the receptors connected to the array it can mean a

⁸ Donald Hebb, 1949 "The organization of behaviour", New York: Wiley

⁹ Walter J. Freeman. 2000 "How brains make up their mind", Columbia University Press

shape and color, a combination of sounds, an irritant or it can have a more abstract meaning.¹⁰

4. **A Neuron soma is a triggered relaxation oscillator**, whereby the oscillation period is multi-rhythmic; various input patterns cause the output pulse width and period to be selected from a limited range, thereby selecting a different population of inter-neurons in each instance. No two instances are ever exactly the same.
5. **Neurons are organized in groups** and each neuron can be a member of many groups. Groups are dynamic; they retain a memory stored within the group, although the memory may change over time. In the past this has been referred to as ‘Population coding of real world events’ but to my knowledge this has never been implemented with a temporal coding system. In practice it appears that each neuron contains a memory segment, which can be combined with other memory segments in other neurons to form a number of different memories. This is not analogous to a binary bit. A memory segment contains more data and of a different nature than a binary bit. A binary bit is a uniform representation of data while a memory segment can have many forms. In simpler terms; a bit is like a Lego block that always fits together with any other block in the same way but can make many combinations. The blocks have only 2 colors. A neuron is a puzzle piece that fits into a memory in a certain way, but changes its color over time. The picture remains, but the puzzle changes over time.
6. **No single neuron or neuron output contains complete complex information.** Complex information, such as a memory, a movement or cognition is encoded in the combined response of a group of neurons. Neurons respond to primitives. Each stimulus can be expressed as a collection of primitives: picture primitives would be, for instance, the dots that combine into a number of horizontal, diagonal and vertical lines, intensity and color information. Unlike computers, brains do not work with ‘pixel’ information. Rather, the horizontal, vertical and diagonal lines, intensity and color are detected in a hierarchy of neurons in which data is reduced to more complex forms. These complex forms consist of the output of a group of neurons. We associate this combined output with a stored pattern for the word “Butterfly” which we have learned in childhood. Prediction within the visual system allows us to fill in the missing bits – in a well known demonstration we see a collection of dots as a dog, or an ink spot as a butterfly.
7. **It is in practical terms impossible to describe the complex evolved brain structure in mathematics.** However, the method that births this complexity can be described. This would see the emergence of a neural computer that grows new circuitry in ‘virgin’ FPGA-like devices, replicating its own neural

¹⁰ Donald Rowe. 2002 “Using experimental data and analysis in EEG modelling” University of Sydney, Behavioural and Brain Sciences, Volume 24, Number 5, 2002

circuits with parameters and connections derived from previously learned knowledge in a Genetic Algorithm.

NeoCortex-1 device (for more details please refer to datasheet)

The NeoCortex-1 device is stackable and contains 140 dynamic synapses and 10 neural processing junctions. The operation of the device is very similar to the way a biological neuron operates; when the input is pulsed the membrane potential is increased or decreased in a 'leaky integrator'. In the biological model the membrane potential increases as vesicles are passed into the synaptic cleft. Over time, the Ca⁺ value increments to a variable maximum value while the integrator value (PSP or membrane potential) declines, e.g. charge 'leaks' out of the integrator. The timing of input signals to each other, the time elapsed since the last input pulse, soma sensitivity and the axon delay are all factors of the transfer function.

