The human vision-processing system is built of numerous complex neural layers that communicate with one another by means of feedforward and feedback neural connections. Via these connections, each neuron frequently makes signals to others at intro-layer or inter-layer locations by broadcasting electrical streams of pulses. Every time a neuron generates a pulse, its addressing information is sensed by a neural junction called synapse, which is temporally connected to a centric sensory line (also known as the bus), where many other neurons are simultaneously competing for the right of way in order to travel further. In such a competition, the general rule is: The recipient neuron at the end of the bus will only listen to neurons that are active when it is active (i.e., the winners are those who have stronger and more consistent signal intensity), and ignore the rest.

A winner-takes-all (WTA) circuit, which identifies the highest signal intensity among multiple inputs, is arguably the most important building block seen in various neural networks, fuzzy control systems, and increasingly often, in integrated image sensors and neuromorphic vision chips that aim to emulate or even outperform—although widely regarded with suspicion—the extremely optic-sensitive coat of the posterior part of the human eye that receives the image produced by the lens; namely, the retina. Once the neuron (also referred to as the cell) with the highest input signal is successfully selected by the WTA circuit, a certain value will be assigned to that winning cell by means of current or voltage, while all other cell’s nominal values will be set to null (i.e., they lose).

Indeed, I’ve found that WTA circuits designed in CMOS (Complementary Metal–Oxide Semiconductor) technology are suitable for the implementation of low-power and high-density neuromorphic chips.

**Subthreshold Properties**

Recalling what our professors said in device physics courses, we find that, when an NMOS transistor’s $V_{gs}$ (the potential between the gate and the source) surpasses $V_{thn}$ (the threshold voltage)—that is, when the value of the effective drain-source voltage $V_{eff}$ exceeds zero—a vertical electric field between the drain and the source is formed, and the device is said to be operating in inversion. When the drain-source voltage ($V_{ds}$) exceeds zero, a horizontal electric field is formed between the drain and the source. The drain-source current ($I_d$) gradually increases with $V_{ds}$ and $V_{gs}$.

(Remember your professor’s water-flowing-through-pipe analogy?)

Once $V_{ds}$ reaches $V_{eff}$, the device’s $I_d$ ~ $V_{ds}$ relationship is described by the following equation (this is
widely known as the square-law I-V relationship),

\[ I_d = \mu_n C_{ox} \frac{W}{L} \left( \frac{V_{ds} - V_{th}}{2} \right)^2 = \mu_n C_{ox} \frac{W V_d^2}{2} \]

\( \mu_n \) is the electron mobility near the silicon surface, \( C_{ox} \) is the gate capacitance per unit area, \( W \) is the gate width, and \( L \) is the effective channel length. If we take into account the channel-length modulation phenomenon (i.e., Early effect) \([1-3]\), then the aforementioned equation should be rewritten to,

\[ I_d = \mu_n C_{ox} \frac{W}{L} \left( \frac{V_{ds} - V_{th}}{2} \right)^2 \left( 1 + \frac{V_{ds}}{V_E} \right) \]

\( V_E \) is the Early voltage and has a general expression as follows \([1-3]\).

\[ V_E = L \frac{\partial V_g}{\partial L} \]

In addition, the Early voltage \( V_E \) is usually employed to characterize a saturated MOS transistor’s small-signal output resistance \( r_{ds} \). That is, \( r_{ds} \approx \frac{V_E}{I_d}[2] \).

The foregoing square-law equations suffice to accurately describe an NMOS transistor operating in saturation, as long as its \( V_{eff} \) satisfies this condition: \( V_{eff} > 4V_T = 4kT/q \). (Sometimes an algebraically convenient alternative, \( V_{ds} > 4V_{th} \), is employed by engineers in their paper-pencil calculations.) When an NMOS transistor’s \( V_{gs} \) is lower than its \( V_{thn} \)—that is, \( V_{eff} \& 0 \)—and yet suffices to sustain a depletion region in the substrate, the transistor is said to be operating in subthreshold. (Typical subthreshold \( V_{eff} \) values range between –100 mV and 0 V.) In such cases, the square law no longer applies.

Due to the presence of a negative \( V_{eff} \) and the resultant diminished vertical electric field underneath the gate, the drift of majority carriers in the NMOS transistor is trivial. As a consequence, \( I_d \) is principally produced by minority-carrier electrons diffusing in the direction of the concentration gradient, which is driven by the horizontal electric field created by \( V_{ds} \). (Minority-carrier electrons will flow from source to drain if \( V_{ds} > 0 \), or from drain to source if \( V_{ds} < 0 \).) In other words, in subthreshold, the depletion current (generated by diffusion) overtakes the inversion current (generated by drift).

Understanding the above, we say that a subthreshold MOS transistor behaves more like a three-terminal diode than a four-terminal field-effect device. In fact, when \( V_{ds} > 0 \), an NMOS transistor operating in subthreshold is almost equivalent to an npn bipolar transistor, where the drain, the substrate and the source act as the collector, the base and the emitter, respectively. (When \( V_{ds} \& 0 \), the drain acts as the emitter and the source as the collector.) \([1-2]\)

Thus, by employing classic device physics formulae appropriate for describing an npn bipolar transistor, we find the general description of \( I_d \) in a subthreshold NMOS, as follows \([1-2]\).

\[ I_d = I_0 \exp \left( \frac{\kappa V_g - V_s}{V_T} \right) \left[ 1 - \exp \left( -\frac{V_{ds}}{V_T} \right) \right] \]

\( I_0 \) is known as the scale current whose value depends mainly on the transistor’s physical geometry \([1]\). The symbol \( \kappa \) stands for the derivative of the surface potential \( \psi_s \) with respect to the gate voltage \( V_{gs} \), which describes a capacitive voltage divider consisting of a gate (or oxide) capacitance
C_{depletion} \) and a depletion capacitance \( C_{\text{depletion}} \) (\( \kappa \) is a constant that takes value between 0.5 and 0.9 [2]).

As we can see, like in an npn bipolar transistor, the current flowing through a subthreshold NMOS transistor varies exponentially with respect to the voltage differences between terminals. Moreover, when \( V_{ds} > 4V_T \), the last term in the foregoing equation is trivial—that is, \( I_d \) is almost independent of \( V_{ds} \)—and the subthreshold transistor is said to be operating in subthreshold saturation. In such a case, given a fixed \( V_{gs} \), the device’s \( I_d \sim V_{ds} \) output curve is almost flat, implying a small dependence shown by \( I_d \) on \( V_{ds} \) whose weight is controlled by the channel-length modulation factor (\( \lambda \)), which is the reciprocal of the Early voltage, \( V_E \) [1-3]. This particular characteristic of a subthreshold transistor is similar to that of an above-threshold one (see Equation (2)).

For describing a saturated subthreshold NMOS transistor, we rewrite Equation (2) as follows,

\[
I_d = I_0 \cdot \exp \left( \frac{\kappa V_g - V_s}{V_T} \right)
\]

Given a fixed source potential (i.e., \( V_s \) is a constant), the NMOS transistor’s transconductance can be expressed by writing the following,

\[
g_m = \frac{\partial I_d}{\partial V_g} = I_0 \cdot \frac{\exp \left( \frac{\kappa V_g - V_s}{V_T} \right)}{\frac{\partial}{\partial V_g} \left( \frac{\kappa V_g - V_s}{V_T} \right)} = I_0 \cdot \exp \left( \frac{\kappa V_g - V_s}{V_T} \right) \cdot \frac{\kappa}{V_T} = I_d \cdot \frac{\kappa}{V_T}
\]

which is similar to the expression of a bipolar transistor’s transconductance.

With the above in mind, we realize that a subthreshold MOS transistor typically has a much larger transconductance-per-unit-current value (\( g_{m}/I_d \)) compared to that of an above-threshold counterpart (assuming both are of identical geometry) [1-3]. As a result, a subthreshold MOS transistor is able to achieve a higher level of sensitivity and a larger voltage gain than its above-threshold counterpart, while both are consuming the same amount of active power. This property is perhaps one of the main reasons why subthreshold MOS transistors are used widely in low-power analog applications such like the CMOS WTA circuit.

Finally, it is instructive to note that a general circuit (e.g., a current mirror) built of subthreshold CMOS transistors is susceptible to imperfections caused by device mismatching, such as an inaccurate current output or a reduced input dynamic range. To a CMOS WTA circuit, these imperfections do not necessarily mean a definite wrong selection, but they do hinder the circuit from tracking its winner adaptively, which more often than not is non-stationary.

**WTA Principles**

The first CMOS winner-take-all (WTA) circuit was reported by [5]. A two-cell version of the circuit is illustrated in Figure 1. It is assumed in this tutorial that corresponding constitutive elements (transistors) of all cells have identical physical geometries.

Let us begin analyzing this circuit by considering the scenario where its two current inputs are identical (\( I_{in1} = I_{in2} = I \)). Transistors \( M_{1a} \) and \( M_{2a} \) act as current sinks that sink the corresponding \( I \) currents into ground. Transistor \( M_{1b} \) and \( M_{2b} \) carry drain currents \( I_{c1} \) and \( I_{c2} \), respectively. Because \( M_{1a} \) and \( M_{2a} \) have identical gate (i.e., \( V_{C1} = V_{C2} = V_c \)) and source potentials while sinking the same amount of current to ground, their drain potentials are identical (i.e., \( V_{out1} = V_{out2} = V_{out} \)). Thus, \( M_{1b} \) and \( M_{2b} \) have identical gate, source and drain potentials, resulting in a couple of identical drain
currents (i.e., $I_{c1} = I_{c2} = I_c$). Assuming both cells have the identical output current mirrors, we have: $I_{out1} = I_{out2} = I_{out}$.

Transistors of each WTA cell in Figure 1 all operate in subthreshold as previously mentioned, offering the possibility of ultra low power dissipation, which are crucial to an integrated neural system. Now, we describe $M_{1a}$ and $M_{2a}$ by employing Equation (4), as follows,

$$I_{out1} = I_{out2} = I_{out}.$$

Transistors of each WTA cell in Figure 1 all operate in subthreshold as previously mentioned, offering the possibility of ultra low power dissipation, which are crucial to an integrated neural system. Now, we describe $M_{1a}$ and $M_{2a}$ by employing Equation (4), as follows,

$$I_{out} = I_{out}.$$

Transistors of each WTA cell in Figure 1 all operate in subthreshold as previously mentioned, offering the possibility of ultra low power dissipation, which are crucial to an integrated neural system. Now, we describe $M_{1a}$ and $M_{2a}$ by employing Equation (4), as follows,

$$I_{out} = I_{out}.$$

Transistors of each WTA cell in Figure 1 all operate in subthreshold as previously mentioned, offering the possibility of ultra low power dissipation, which are crucial to an integrated neural system. Now, we describe $M_{1a}$ and $M_{2a}$ by employing Equation (4), as follows,

$$I_{out} = I_{out}.$$

Transistors of each WTA cell in Figure 1 all operate in subthreshold as previously mentioned, offering the possibility of ultra low power dissipation, which are crucial to an integrated neural system. Now, we describe $M_{1a}$ and $M_{2a}$ by employing Equation (4), as follows,

$$I_{out} = I_{out}.$$

Transistors of each WTA cell in Figure 1 all operate in subthreshold as previously mentioned, offering the possibility of ultra low power dissipation, which are crucial to an integrated neural system. Now, we describe $M_{1a}$ and $M_{2a}$ by employing Equation (4), as follows,

$$I_{out} = I_{out}.$$

Transistors of each WTA cell in Figure 1 all operate in subthreshold as previously mentioned, offering the possibility of ultra low power dissipation, which are crucial to an integrated neural system. Now, we describe $M_{1a}$ and $M_{2a}$ by employing Equation (4), as follows,

$$I_{out} = I_{out}.$$
From the above, we see that the change in $I_{c1}$ due to $\Delta I$ is indeed given by,

\[ E_{\text{gain}} = \frac{\kappa}{V_T} \Delta I = \frac{I_{\text{bias}} \cdot \kappa}{V_T} \Delta I = \left( \frac{\kappa \cdot V_T}{I_{\text{bias}}} \right) \Delta I \]

The parenthesized term in the foregoing equation stands for the gain factor that describes how sensitive the cell is with respect to the input difference. By assigning typical values to these parameters—$\kappa = 0.5$, $V_E = 50$ V, and $V_T = 26$ mV—we get a gain of about 962 or equivalently, 60 dB.

Due to such a large gain factor, a very small $\Delta I$ will be sufficient to cause a fairly significant increase in $I_{c1}$ (and this increase is to be supplied by the losing cell’s current source). For instance, given the aforementioned parameter values and that $\Delta I = 0.001 I_{\text{bias}}$, a $\Delta I_{c1}$ as large as 0.962 results according to Equation (11).

Finally, from Equation (3), we realize that the longer the channel of transistor $M_{1a}$ (or $M_{2a}$), the larger the Early voltage and in turn the more sensitive the WTA circuit will be (in response to the input difference)—that is, the WTA circuit will demonstrate a steeper winning/losing response.

In concluding this section, we say that the circuit shown in Figure 1 is capable of selecting the cell with the highest input current, regardless of the extent to which the cell surpasses its competitors. In addition, it can be shown that the properties mentioned above apply to an n-cell (n > 2) WTA circuit [5].

**WTA with Local Inhibitory Decoupling**

Besides the circuit shown in Figure 1, a WTA circuit that contains local inhibitory decoupling (quite often the term local inhibitory coupling is used instead) was reported by [5]. The purpose of the local inhibitory decoupling feature is to contain a winner’s influence within a prescribed spatial range—that is, in such a WTA circuit, a winning cell inhibits its neighboring cells but not cells that are sufficiently distant (or, decoupled) from it—resulting in a fork-like spatial impulse response. (In a sense, it has the shape of a Chinese character “SHAN”, which means “the mountain”). As reported by [5], resistors (which deter current) can be used in a WTA circuit to realize the inhibitory coupling functionality, creating buffers between $V_c$ nodes such that every local winner is allowed to take some, but not all the bias currents.

A slightly modified version of the said circuit is illustrated in Figure 2, where aforementioned resistors are substituted with NMOS transistors operating in subthreshold saturation (shown as $M_{1c}$ and $M_{2c}$ in Figure 2), for achieving higher area density and lower thermal dissipation. In this circuit, the gate potentials of $M_{1c}$ and $M_{2c}$ are both controlled by the Inhibition voltage, $V_{\text{inhi}}$. By employing Equation (5), we write the expression of each drain current as follows,

\[
\begin{align*}
I_{d1c} &= I_0 \cdot \exp \left( \frac{\kappa V_{\text{inhi}} - V_{C1}}{V_T} \right), \\
I_{d2c} &= I_0 \cdot \exp \left( \frac{\kappa V_{\text{inhi}} - V_{C2}}{V_T} \right)
\end{align*}
\]

As we can see, the current flowing through $M_{1c}$ ($M_{2c}$) is exponentially dependent upon $V_{\text{inhi}}$. The
higher the $V_{\text{inh}}$ voltage, the more current flows toward the winner’s $V_C$ node through the local inhibitory decoupling transistors. If $V_{\text{inh}}$ reaches the highest possible potential on chip (say, $V_{dd}$), then the drain and the source of $M_{ic}$ are literally shorted together, resulting in zero degree of local inhibitory decoupling (i.e., identical to the circuit shown in Figure 1).

On the other hand, if $V_{\text{inh}}$ is sufficiently decreased such that $M_{ic}$ is effectively turned off, then the maximum degree of decoupling is achieved and as a result, any cell can be selected as the winner, independent of its neighbors. In a word, by adjusting the value of $V_{\text{inh}}$, we can control the amount of inhibitory current flowing laterally, which in turn determines how many neighboring cells a local winner is allowed to suppress (inhibit).

The aforementioned circuit idea is originated from a novel configuration called pseudo-conductance current divider (diffusor) [4], which basically exploits the similarity between a subthreshold MOS transistor and a bipolar transistor: The current flowing in a subthreshold MOS transistor, like that in a bipolar transistor, can be divided into forward and reverse components; that is, $I_d = I_f - I_r$. With this in mind, we revisit the decoupling transistor $M_{ic}$ and obtain the following:

\[
\begin{align*}
I_f &= I_0 \cdot \exp \left( \frac{\kappa V_{\text{inh}} - V_{C1}}{V_T} \right) \\
I_r &= I_0 \cdot \exp \left( \frac{\kappa V_{\text{inh}} - V_{C2}}{V_T} \right), \quad \text{and thus} \\
I_{d1c} &= I_f - I_r = I_0 \cdot \exp \left( \frac{\kappa V_{\text{inh}} - V_{C1}}{V_T} \right) \left( 1 - \exp \left( \frac{V_{C1} - V_{C2}}{V_T} \right) \right)
\end{align*}
\]

Assuming that the value of $(V_{C1} - V_{C2})$ is sufficiently small compared to $V_T$, we can rewrite the foregoing expression of $I_{d1c}$ back to Equation (12). In addition, assuming that the current-source transistor of the winning Cell 1 is operating in subthreshold, we find that,

\[
\text{Equation 14} \quad \frac{I_{d1c}}{I_{d2c}} \approx \exp \left( \frac{\kappa (V_{\text{inh}} - V_{\text{nd}})}{V_T} \right)
\]

The larger this ratio is, the less effective the decoupling will be.

**WTA with Local Excitatory Coupling**

The same methodology of current diffusor networks (as mentioned) can be adopted to analyze a third type of WTA circuit, which contains both local inhibitory decoupling and local excitatory coupling[6]. The purpose of the local excitatory coupling feature is to slightly stimulate each locally inhibited cell, thereby increasing its intrinsic output value (hence the term excitatory). The outcome is a smoother overall spatial impulse response.

Intuitively speaking, while local inhibitory decoupling creates fork-like output characteristics, local excitatory coupling introduces an output curve that looks more like a witch’s hat—close your eyes and picture one of Harry Potter’s magical teachers in the movie—which is smoother compared to the fork but still has rather steep slopes.

A modified version of this circuit is shown in Figure 3, where the local excitatory coupling transistor (e.g., $M_{id}$) is connected between nodes $V_{out1}$ and $V_{out2}$, rather than between $V_{F1}$ and $V_{F2}$, facilitating the
use of NMOS excitatory transistors only (otherwise, $V_{\text{exc}}$ would need to be set at values higher than $V_{\text{dd}}$[7]).

As we can see in Figure 3, unlike the previous case of local inhibitory decoupling, where $M_{1c}$ and the two current-source transistors form a current diffusor network, here in Cell 1, transistors $M_{1d}, M_{1a}$, and $M_{2a}$ form a current diffusor network. Consequently, the extent to which the local winning effect ripples laterally depends very little on the bias current ($I_{\text{bias}}$); rather, it is closely related to how much current actually passes through $M_{1a}$ and $M_{2a}$ to ground. This aspect translates to a relationship between the rate of spatial smoothing (also called spatial filtering) and voltage potentials $V_{\text{exc}}, V_{C1}$, and $V_{C2}$. Revisiting Equation (13), we find the following relations with respect to $M_{1d}, M_{1a}$, and $M_{2a}$,

\[
\begin{align*}
I_f &= I_0 \cdot \exp \left( \frac{kV_{\text{exc}} - V_{\text{out}1}}{V_T} \right), \\
I_r &= I_0 \cdot \exp \left( \frac{kV_{\text{exc}} - V_{\text{out}2}}{V_T} \right), \\
I_{1a} &\approx I_0 \cdot \exp \left( \frac{kV_{C1} - V_{\text{out}1}}{V_T} \right), \\
I_{2a} &\approx I_0 \cdot \exp \left( \frac{kV_{C2} - V_{\text{out}2}}{V_T} \right).
\end{align*}
\]

Equation 15

Assuming that Cell 1 is the local winner and that $V_{C1} = V_{C2} = V_{C}$, we obtain,

\[
\frac{I_{1a}}{I_{2a}} = \frac{I_f - I_r}{I_{1a} - I_{2a}} \approx \exp \left[ \frac{k(V_{\text{exc}} - V_{C})}{V_T} \right]
\]

Equation 16

The rightmost term in the preceding equation is called the WTA circuit’s space constant ($\lambda$), and it is useful for describing the trend of signal loss with respect to the distance from the local winner. Now, let us add up all excitatory currents, by applying Equation (16) recursively across the entire spatial network. After utilizing the famous Euler’s formula, we arrive at the following conclusion,

\[
I_d(n) = V_d(0) \cdot \lambda^n
\]

Equation 17

From the two preceding equations, we see that the space constant $\lambda$ is controlled by the value of $(V_{\text{exc}} - V_{C})$ exponentially, and that $\lambda$ shall be smaller than unity. As mentioned, the purpose of local excitatory coupling is to prevent the neighboring inhibited cells from dying out (too fast), and thus, for a WTA circuit with excitatory coupling to work properly, $V_{\text{exc}}$ has to be set at smaller than $V_{C}$. Additionally, given a fixed $V_{C}$ value, the larger the $V_{\text{exc}}$ voltage, the slower the rate at which the inhibited cells lose, meaning that the circuit will have a wider and smoother spatial impulse response.

This above quantitative result matches that of an intuitive observation, which is described as follows: If $V_{\text{exc}}$ is so large that all excitatory transistors are effectively shorted, then all output nodes ($V_{\text{out}1}$) share the same potential, thereby creating a situation where neither a winner nor a loser exists, which eventually defeats the purpose of the WTA circuit. On the contrary, if $V_{\text{exc}}$ is sufficiently decreased such that all excitatory transistors are effectively turned off, then the circuit of Figure 2 results.

**Hysteretic WTA with Local Excitatory Feedback**
Many CMOS integrated circuit designers are familiar with the technique of utilizing unsymmetrical decision circuit to add hysteretic properties into a CMOS voltage comparator [9]. In a typical CMOS WTA circuit—which in essence is a reset-able multiple-input current comparator—similar hysteretic properties can be realized by means of local excitatory feedbacks[10-12]. More often than not, local excitatory feedbacks are used in combination with the aforementioned current diffusing (or distributing) mechanism [6-8] to create an “all-in-one” sort of WTA circuit topology [13]. An example of such circuits is illustrated in Figure 4, which was originally reported by [13] and further discussed by [8].

Looking into Figure 4, we identify two aspects that are not employed in any previous configuration: Diode-source degeneration implemented by means of transistors $M_{1S}$ and $M_{2S}$, and local excitatory feedback implemented by means of transistors $M_{1F}$ and $M_{2F}$. The former is a fairly straightforward setup, for it is well known that connecting a diode-connected transistor to a common-source single-stage amplifier increases the amplifier’s output impedance; that is, the ratio ($\Delta V_{out}/\Delta I$) is increased due to $M_{1S}$ ($M_{2S}$), thereby producing a narrower losing response [5]. The latter is responsible for introducing a hysteretic behavior to the WTA circuit.

Specifically, if Cell 1 becomes the winner, then a large amount of current will flow through $M_{1b}$. Due to the PMOS current mirror that includes $M_{1p}$, a small portion of the large $M_{1b}$ current, whose exact amount ($I_{fb}$) can be adjusted by sizing the PMOS transistors, is positively fed back to the output node $V_{out1}$ (as the arrow shows). In other words, $I_{fb}$ is added into the original input current $I_{in1}$. As a result, during the next selection, the condition for Cell 1 to be de-selected will be: There must be a potential winner whose input current is larger than ($I_{in1}+I_{fb}$). That is, $I_{fb}$ is the hysteretic current.

Similar to the case of hysteretic voltage comparator, hysteretic current $I_{fb}$ effectively spares a current-mode WTA from making erroneous selections under the influence of noise, device mismatch, or dc offset.

In addition, thanks to the existing local excitatory coupling configurations (reference to $M_{1d}$ and $M_{2d}$ in Figure 4), the hysteretic current $I_{fb}$ is distributed to the neighboring cells’ input nodes. However, knowing that to any cell an increased input current generally means a higher $V_{C}$, we find based on Equation (16) and Equation (17) that in a sense, $I_{fb}$ works against local excitatory coupling by distributing hysteretic protection to the neighboring cells, which helps the existing local winner maintain its winning status as it shifts from one spatial position to another. Therefore, there is a trade-off to balance by appropriately sizing the transistors (especially the PMOS current mirror and the coupling transistor). This particular property is instrumental to applications that require for constantly tracking the strongest input signal, which is non-stationary [13-14].

More WTA Circuits

Quite a few novel CMOS WTA circuits have been reported in the literature since the publication of [5], most of which are endowed with their respective strengths and weaknesses (in speed, resolution, power, or density).

Figure 5 illustrates a practical CMOS circuit that may function as either a regular WTA (i.e., there is only one winner) or a compromised WTA (i.e., a soft-max configuration where there can be multiple winners simultaneously), depending on the adjustable voltage $V_A$[15]. Specifically, when $V_A$ is smaller than all output voltages $V_{outi}$, the circuit performs soft-max computations; otherwise, the diode-
connected transistors $M_{1g}$ and $M_{2g}$ are effectively turned off, and thus the circuit performs WTA computations.

Intuitively speaking, connecting a diode-connected transistor to node $V_{out}$ is similar to paralleling $M_{1a}$ with a physical resistor—remember that we regarded $M_{1a}$ as a resistor $r_{ds}$ when analyzing the first WTA circuit—and in effect, that changes $V_{out}$ from a high-impedance node (as in WTA configurations) into a low-impedance node (as in soft-max configurations). Given a fixed conductance of $M_{1b}$, such a decrease in nodal impedance hinders the cell with the strongest input from actually winning over the others, thereby alleviating the likelihood of an all-time single-cell dictatorship. For brevity, quantitative circuit analysis is not included here, and the reader is referred to the said paper [15] for more details.

Another interesting WTA topology is illustrated in Figure 6[16]. Here, only one cell is shown for simplicity. The underlying idea is similar to that of [11] or [12]: that is, not only local excitatory feedback (through $M_{9}$, $M_{6}$, and $M_{5}$) is adopted to add hysteresis to the circuit, but a so-called local inhibitory feedback (through transistors $M_{3}$, $M_{7}$, and $M_{8}$) is employed here. In essence, local inhibitory feedback is a form of intro-cell local inhibitory decoupling (as opposed to the aforementioned inter-cell one) with a dependency on the cell’s output voltage. When $V_{reset}$ is set to “1”, both transistors $M_{ij}$ and $M_{ki}$ are effectively turned off, thereby cancelling out the feedback current parameters (excitatory and inhibitory) from the previous selection. During the next cycle, $V_{reset}$ is set to “0” till a winner is selected.

Besides current-mode WTA circuits, additional works on voltage-mode WTA design have been reported [17-19]. In particular, the authors of [19] explored the possibility of eliminating the circuit’s dependence upon device matching characteristic, by employing an inverter-based current comparator as the core for each WTA cell. To probe further, perhaps one should investigate the feasibility of substituting the said inverter with a CMOS Schmitt trigger, which adds hysteretic properties into the circuit and enhances the overall robustness.

**Loser-Take-All (LTA)**

More often than not, it is desirable to select and track the loser (i.e., the cell with the smallest input) rather than the winner in a neural network. By subtracting the winner from a fixed reference, we can get the loser. However, in the analog domain, subtraction and addition are two functionalities subject to circuit-implementation problems, such as loss of accuracy and reduction of input/output dynamic range.

A CMOS loser-take-all (LTA) circuit that does not require for implementing subtraction in the analog domain is shown in Figure 7[20]. The circuit operates as follows. Assuming that Cell 1 is the potential loser, we find that $V_{out}$ is the smallest output, and hence $M_{1b}$ is turned off. As a result, the PMOS transistor $M_{1m}$ is turned on, and the bias current supplied by the global bias cell ($I_{bias}$) flows through $M_{1m}$ and $M_{in}$ to ground. Transistor $M_{1m}$’s non-zero drain-source voltage serves as the digital output pulse, broadcasting across the network that Cell 1 has been selected as the loser. In addition, the loser’s input current $I_{inl}$ is copied to the signal path composed of transistor $M_{y}$, $M_{a}$, and $M_{s}$, which is read out in the analog domain as the minimum input value.

In conclusion to create CMOS winner-take-all (WTA) circuits one must understand the essential properties of subthreshold MOS transistor as well as be familiar with the various CMOS WTA and loser-take-all (LTA) circuits.