Analogue building blocks for neural-inspired circuits

Steve Hall
University of Liverpool, s.hall@liv.ac.uk

Liam McDaid
University of Ulster, lj.mcdaid@ulster.ac.uk
Some facts about the brain as a PC...

• The brain has ~100 billion neurons \((10^{11})\) – about 30µm large
  – Neuron Fan-in \(\sim 10^3 - 10^4\) (logic gates 2-4!)
  – complex dynamics - includes several time constants,
  – maintains a more complex internal state
  – output is a time-series of action potentials
    or ‘spikes’ - no information in amplitude!

• Massively parallel in nature
  – Typical \(10^{15}\) interconnections
  – Total computation rate of about \(10^{16}\) complex operations /sec (cf 10 P-FLOPs)

• Millisecond time frame of ‘events’

• Low level function: ‘reasonably well understood’...

• High level function.........................????????
Some other brains

- A fly (1 grain of sugar a day to feed it!): 250 k neurons
- Honeybee (fantastic navigator!): 1 million neurons
- Rat (pretty smart animal): 55 million neurons

- But how do the following work:
  - the arithmetic
  - Fault-tolerance
  - The parallelism (beat Moore’s Law hands down)

This is the inspiration!
But must find a simpler, scaleable, low power approach
Synapses and neurons

Spike-timing dependent plasticity

If spike, $t_1$ causes neuron, $N$ to fire ($t_2 - t_1$ small).

Weight $W_1$ may be increased

and $W_2$ etc decreased

STDP learning rule
Motivation

Create building blocks that can emulate biological functionality

Implement in mixed signal CMOS (cheap!)

Assess layout / scalability / systems functionality

Circuits that can learn!
  - Plasticity / decision circuits (STDP) / FG weight storage

Build large, useful electronic systems
learn more about ‘brain computation’ .....
Circuit Challenges

- Store and update weights
- Detect timing \( (t_2 - t_1) \)
- Axonal delay
- Low power operation
- Scale to VLSI
- Learn!
Dynamic synapse

\[ V_{\text{pres}} \rightarrow \text{on} \]
Charge sharing
\[ S \text{ of } M2 \text{ increase} \]
\[ \rightarrow M2 \text{ clamped \textit{`off'}} \]
Transient \( i(t) \), mirrored in M5

\[ \begin{align*}
V_{\text{pres}} &= 0 \text{ V} \\
V_{\text{pres}} &= V_{\text{DD}} \\
V_{p} &< V_{T} \\
V_{W} &
\end{align*} \]

$V_{IN} = Q_W / C_{IN}$
Post-synaptic potentials

\[ V_{\text{DD}} \]

\[ V_{\text{IN}} \]

\[ V_{\text{PRES}} \]

\[ V_{\text{DM2}} \]

\[ V_{\text{W}} \]

\[ V_{\text{DM1}} \]

\[ V_{\text{P}} \]

\[ V_{\text{LEAK}} \]

\[ V_{\text{PSP}} \]

\[ C_{\text{PSP}} \]

\[ \text{ISI} = 250 \text{us} \]

\[ \text{ISI} = 200 \text{us} \]

\[ \text{ISI} = 200 \text{us} \]

\[ \text{ISI} = 150 \text{us} \]

\[ \text{ISI} = 20 \text{us} \]

\[ \text{ISI} = 150 \text{us} \]

\[ \text{ISI} = 100 \text{us} \]

\[ T \text{ime (s)} \]

\[ V_{\text{PSP}} (V) \]
Fan-in: theory
Consider transients of capacitive nodes

\[ \Delta V_{IN}(t) = \frac{Q_W}{C_{IN}(n)} = \frac{C_{OX}(V_W - V_{Tn})}{C_{IN}(n)} \]

\[ C_{PSP} \frac{dV_{PSP}(t)}{dt} = I_{Op} \exp\left(\frac{V_{IN}(t)}{m_PV_t}\right) - I_{M6} \]

Rise time

\[ \tau_R = \left[ \frac{I_{Op}}{m_PV_tC_{IN}} \right]^{-1}\left\{ \frac{I_{On}}{I_{Op}} \exp\left(\frac{V_{LEAK}}{m_nV_t}\right) - \exp\left(\frac{-V_{IN}(0)}{m_PV_t}\right) \right\} \]

Post-synaptic potential

\[ V_{PSP}(t) = m_PV_t \frac{C_{IN}}{C_{PSP}} \left[ 1 + \frac{I_{Op}}{m_PV_tC_{IN}} \exp\left(\frac{V_{IN}(0)}{m_PV_t}\right) \right] - \frac{I_{M6}}{C_{PSP}} t \]

Fan-in

\[ V_{PSPMAX} = m_p V_t \frac{C_{IN}(n)}{C_{PSP}} \ell n \left[ 1 + \frac{I_{OP}}{m_p V_t C_{IN}(n)} \exp \left( \frac{V_{IN}'(0)}{m_p V_t} \tau_R \right) \right] - \frac{I_{M6}}{C_{PSP}} \tau_R \]

Conclusion: Fan-in intrinsic limit > 10^5!
Practical limit is set by layout / interconnect

Compact decision circuits (STDP)

**Weight Increase, WI, Circuit Block, Output Buffers and SIFGNVM Device**

- **Pass transistors gated by** $V_{\text{pre}}$, $V_{\text{post}}$
- **Charge node X**
- **Decay via sub-Vth MOST**
- **Sets plasticity ‘window’**

*Smith et al, Neurocomputing, v124, p 210 (2014)*
How it works: WI Block Operation Pre-Post Spiking Event

- When a presynaptic spike occurs ($V_{\text{Pre}}$)
  - $V_1$ is pulled up to $3V - V_{\text{TMpre}}(V_1)$, $C_1$ charges via $M_{\text{pre}}$
  - $C_1$ Slowly discharges via sub-threshold $M_{\text{leak}}$
  - $V_{\text{post}}$ triggers the sample/hold as some time, $t$ after $V_{\text{pre}}$
Axonal delay

- $M_1$ operates subthreshold
- Slow charging of $C$
- $V_N$ rises and inverters turn on
- Tune delay with $V_{LEAK}$

\[
\tau_d = \frac{C}{I_0 \exp \left( \frac{V_{GSM1}}{mV_{th}} \right) (V_{TRIG} - V_{N0})}
\]

Dowrick et al, Neurocomputing, 2012
http://dx.doi.org/10.1016/j.neucom.2012.12.004
Pulse burst creation

Add feedback (M₁₀)
Define pulse trains

Integrate axon delays (A) into paths

Dowrick et al, Neurocomputing, 2012
http://dx.doi.org/10.1016/j.neucom.2012.12.004,
Scaling

- Two solutions: sum voltages or sum currents
Scaleability: easier to sum currents

Transmit voltage steps and re-create spikes for long interconnect

But added complexity!
Scaling: circuit issues

Large synapse fan-out problem:
non uniform spike inputs due to parasitics
non-linearities occur in currents

Hope it all comes out in the wash!
Nature is messy as well
Neurons with excitatory and inhibitory synapses

\[ \sum_{m,n} I_{psc_{ex}} - \sum_{p,q} I_{psc_{inh}} - I_{th} \geq 0 \]
Programmable weights

• Analog weight
  – Good: Continuous weight value, compact analog storage circuit
  – Bad: Inaccurate, require bias reference circuit and complex control circuit
    for high resolution, also require high voltage rail and undocumented
  – technology feature

• Digital weight
  – Good: accurate, mature digital memory technology, easy to program
  – Bad: discrete quantitative weight, require more space
Programmable weight

Synapse Block
Embrace: an alternative approach

- Network-on-chip address the issues of scalability and connectivity between components.
- Low-area/power spiking neuron cells with associated training provides neural computing capability.

- 2-dimensional array of interconnected neural tiles + I/O blocks.
- Neural tiles connected in North, East, South and West.
- Tile can be programmed to realise neuron-level functions.


*Slide courtesy of Jim Harkin*
Evaluation

• Learning in software (calculate weight values)
  • Fit the experimental synapse results
• Solve benchmark problems
  • Wisconsin breast cancer (WBC) dataset
  • IRIS dataset
• Temporally encoded input values

SNN architecture: IRIS dataset

SNN architecture: WBC dataset

Circuits fabricated in AMS 0.35, mixed signal CMOS
Astrocytes

Study transport within astrocyte process and between neuron/astrocyte

Endocannabinoid Mediated Self-Repair

10 synapses

Remains undamaged

Damage a fraction of these synapses

Astrocytes mediate self-repair

- Astrocyte ‘forces’ synapses to ‘work harder’
- Opens up STDP window – restarts learning

Wade, McDaid et al, Frontiers in computational neuroscience, v6, Article 76 (2012)
What we learnt..

- Can build compact analogue circuits that emulate aspects of biology with a degree of success (better than in software? – potentially much much faster)
- Getting them to learn is another matter..
  - Need feedback
  - Weight update
  - Starts to get very complicated…
- A lot of redundancy once the circuit has ‘learnt’
- Scaling soon results in a huge amount of interconnect

Need software/hardware combination – learning in software
Still some way to go before....
Thanks to

J Harkin  (jg.harkin@ulster.ac.uk)

T Dowrick (PhD)
A Smith (PhD)
S Chen (PhD)
S Zhang (post-doc)
A Ghani (post-doc)

**Funding:** EPSRC, NAP, EPSRC-DTA awards, Dorothy Hodgkin scholarship