E-I Balanced Point Attractor Network on Neuromorphic Chips

Neuromorphic Intelligence Lecture Project Report

Institute of Neuroinformatics
University of Zurich & ETH Zurich

Wenjie He & Junren Chen

Supervision

Giacomo Indiveri
Elisa Donati
Jingyue Zhao

June 2021

Contents

Chapter 1

Introduction of E-I balanced Network and Point Attractor Network

1.1 E-I balanced Network

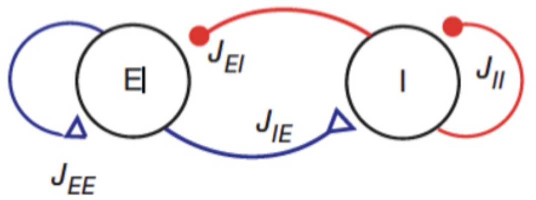
Working memory in the brain can be interpreted as a persistent brain neuron activity even in the absence of an external stimulus.
It is found that population connection with positive and negative feedback or hybrid connection can lead to a persistent activity with a given pulse input into the neuron population.
The neuron populations with excitatory connections (both to itself and to other populations) are named as E neuron populations which will excite the following connected neuron, and the neuron populations with inhibitory connections are named as I neuron populations which will inhibit the following connected neuron. A model that is used for explain the persistent activities is shown in figure 1.1.
A self-positive excitation of E population is considered as positive feedback (J_{EE}) and the negative excitation of E population is considered as negative feedback(J_{II}). The interconnection between E and I population is considered as shown in the figure 1.1 is considered as derivative feedback (J_{EI})[2].
In this model, with pure positive feedback, the network can reach a persistent state, but it requires very careful parameter tuning and is not stable when external perturbance appears.
However, it is found that with a derivative feedback connection, the network can show the persistent activity by simply tuning the parameters and can be stable when external perturbance appears.
This kind of network can be considered as a E-I balanced network.

1.2 Point Attractor Network

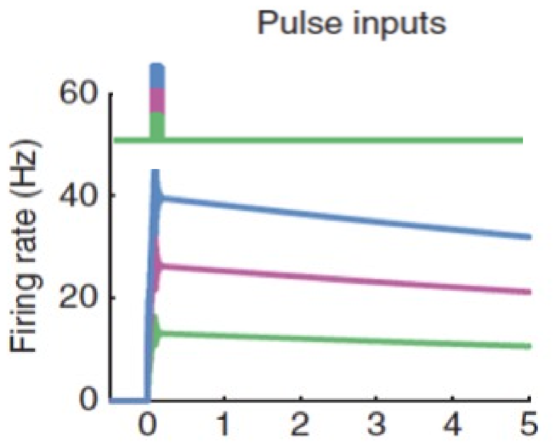
Besides the brief introduction of the theory of E-I balanced neural network, a VLSI neuromorphic chip is developed for emulating the neuronal behaviours. By connecting the neuron populations on this VLSI device, a E-I balanced network is realized which can implement a function as attracting the E population neurons to a persistent active state from a silent state by input excitatory stimuli. This active state is called an "attractor" state, and keep active even the external stimuli is removed. Since the network can stay at two stable states (one silent and one active),it is named as a point attractor network. The point attractor network is self-correcting (which means it is robust to perturbation) and self-sustaining even without external stimulation [1].

1.2.1 Network structure

The architecture of the point attractor, which was implemented on a VLSI neuromorphic chip point attractor network is shown in figure 1.2 [1].



(a) The neuron population connection structure where the E population can realise a persistent activity.



(b) Persistent activity behaviour: given a pulse input, the neuron population will remain it's firing state for a longer period.

Figure 1.1: The derivative network structure[2] and persistent activity of a neuron population

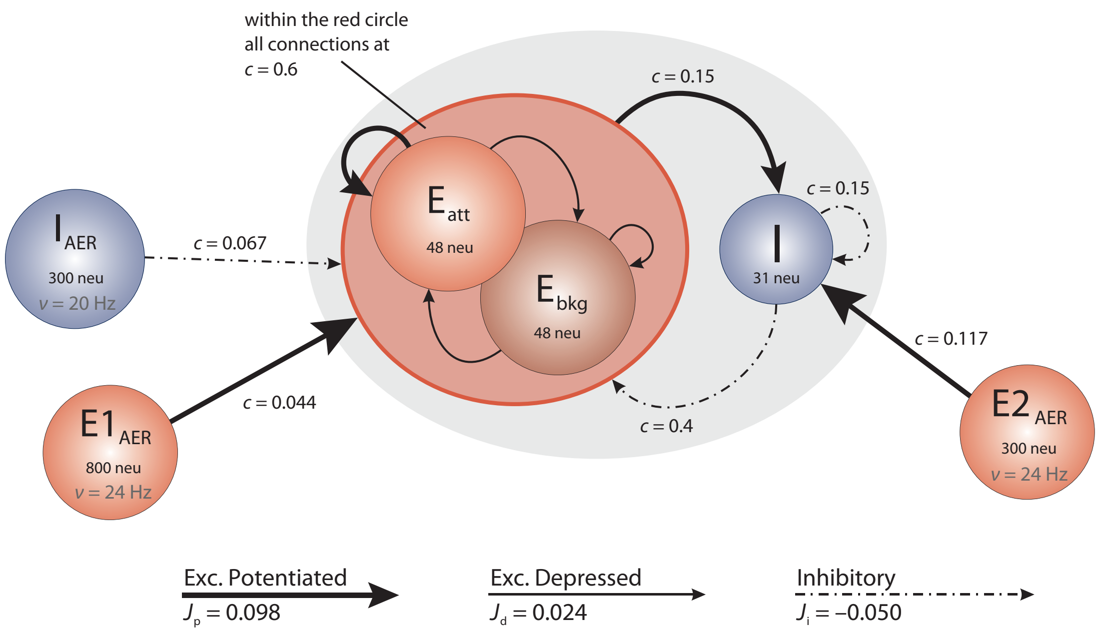


Figure 1.2: Network structure implemented on a VLSI chip to realize a point attractor network.

The excitatory population is divided into E_{att} and E_{bkg} populations, consisting of 48 neurons individually. The I population consists 31 neurons.
The E_{att} neurons are self-connected with excitatory potentiated synapses and E_{bkg} neurons are self-connected with excitatory depressed synapses. E_{bkg} and E_{att} populations are inter-connected with a excitatory depressed synapse. The excitatory potentiated synapse has a higher weight compared with excitatory depressed one. The E population connection probability is set as $c = 0.6$.
The E population feeds forward its firing activity to I population with excitatory potentiated synapses of a connection probability $c = 0.15$. The I population feeds forward its firing activity to E population with inhibitory synapses of a connection probability $c = 0.4$. The I population neurons are self-connected with inhibitory synapses of a probability $c = 0.15$.

Besides the neuron populations, there are three input Poisson spike generators named as $E1_{AER}$, $E1_{AER}$ and I_{AER}. These spike generators consists of 800, 300 and 300 neurons respectively and each one feeds in firing rate of 24Hz, 24Hz and 20Hz respectively.
The $E1_{AER}$ and I_{AER} both feed into E neuron populations on chip with a probability of $c = 0.044$ and $c = 0.067$ respectively. The $E1_{AER}$ helps excite the E population on chip and I_{AER} helps inhibit the E population. The $E2_{AER}$ feeds into I neuron population on chip with a probability of $c = 0.117$. The $E2_{AER}$ helps excite the I population.

1.2.2 The behaviour of point attractor network

The input stimuli and neuron population firing behaviour is shown in figure 1.3. It is discovered that the excitatory population has two states: "low" and "high" state where the neuron population fires persistently at a low and high rate respectively.

The input profile of E population is shown in the lower part of the figure. At 0.5s, when a small excitation (34Hz input) is given to the E population, the firing rate of E population increases accordingly. But the state of the network is not changed and keeps at a "low" firing state. When the weak input stimuli is retreated, E population goes back to relative silent state.

When high excitations kick in (67Hz, 84Hz or 115Hz), the network changed into a "high" state where the E population fires at a rate higher than 100Hz. When the strong input stimuli is retreated, the E population keeps at the high state where E_{att} remains to fire at a rate higher than 100Hz and E_{bkg} fires at a rate higher than 20Hz.

The high state of E population is stable even when a perturbation kick in. At 3s, when a small inhibition fed in, the E population firing rate fluctuates but after the small input inhibition, the E population goes back to a firing rate at high state. Only when a strong inhibition kicks in, the E population will change the state back to "low".
It is suggested that the recurrent feedback connection of E population and balance between E-I population is the reason behind the maintaining of population high firing rate state.

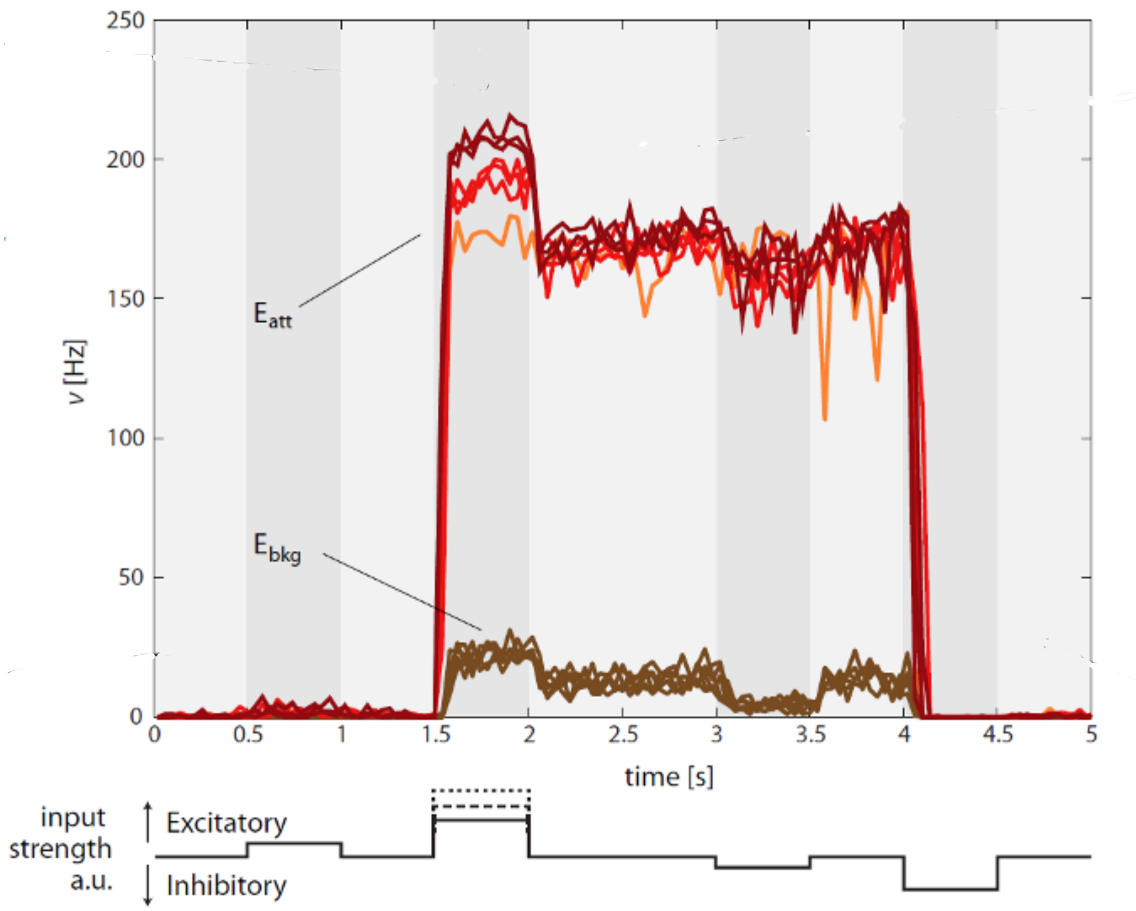


Figure 1.3: Point attractor network's input stimuli and neuron population firing rate [1]. The attractor state is obtained after a strong stimuli at 1.5s.

Chapter 2

Point Attractor Network on a Neuromorphic Chip

2.1 Bias (parameters) on chip

2.1.1 Linear properties of bias settings

Each core on the chip share the same synapse properties. Before tuning the parameters to obtain a point attractor behaviour, the relation between the bias we set and the bias the board applied is investigated.

Each bias contains two values, one coarse value ranging from 0-7 and a fine value ranging from 0-255. The given coarse value and fine value will be translated into a linear value which will then be applied on the chip.

The relationship between the coarse value and fine value is shown in figure 2.1. Figure 2.1a shows that in each coarse values settings with varying fine values, the translated linear value all started from 0, the higher the coarse value, the higher linear value range the bias can reach by tuning the fine values. Figure 2.1b shows all the translated bias values in one coarse set are linear to the fine values with different slopes. A log scale plot as shown in figure 2.1c is made to make the relationship between different coarse sets more readable.

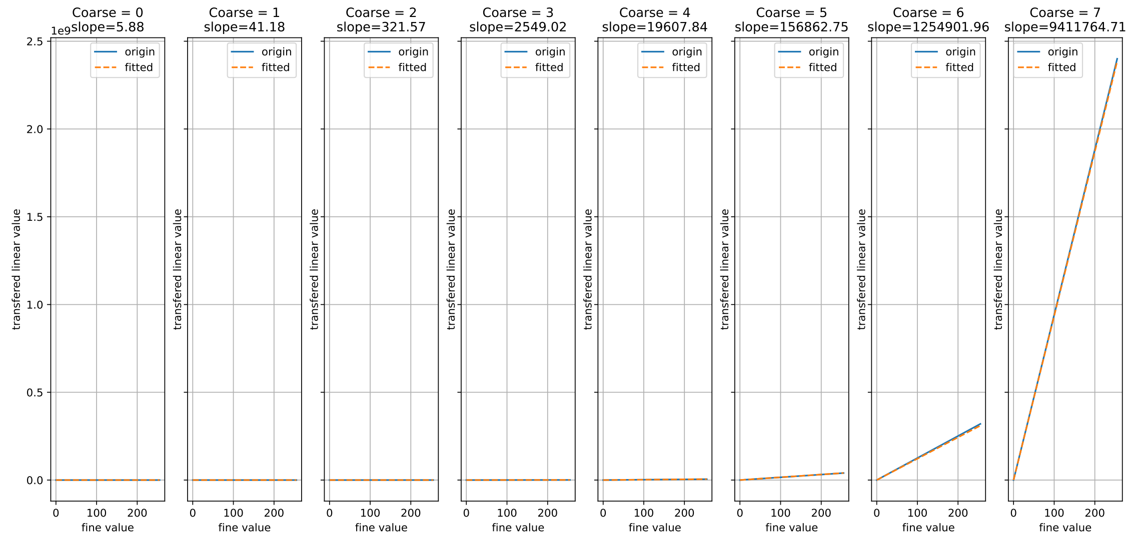
The slopes of each coarse set are measured and listed in table 2.1. After taking the logarithm of the slope, a linear relationship is found as shown in figure 2.2.
Therefore, the coarse and fine value bias translation is approximated by a function $y = 5.88e^{2.04c}f$, where c represents coarse value and f represents fine value). The fitted slope and corresponding bias settings are plotted into the figures and find a good fit between the measurement of board 020 and the curve fitting.

2.1.2 Synapse properties on chip

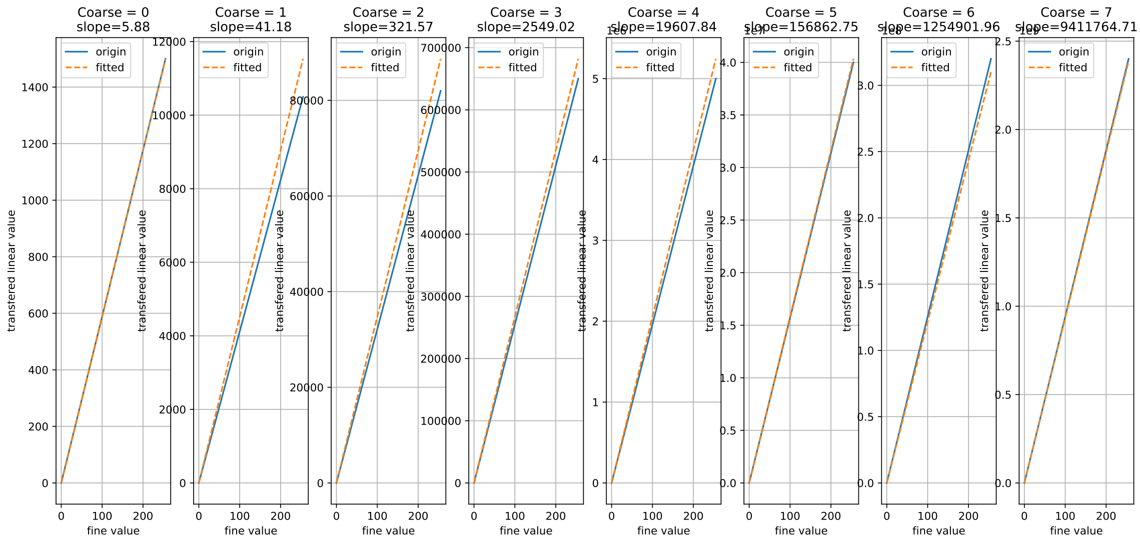
As shown in figure 2.3, given the same parameter settings for AMPA and NMDA as shown in table **??**, the firing rate of the neuron seems to be identical, therefore, different from the brian2 software

coarse	0	1	2	3	4	5	6	7
slope	5.88	41.18	321.57	2549.02	19607.84	156862.75	1254901.96	9411764.71

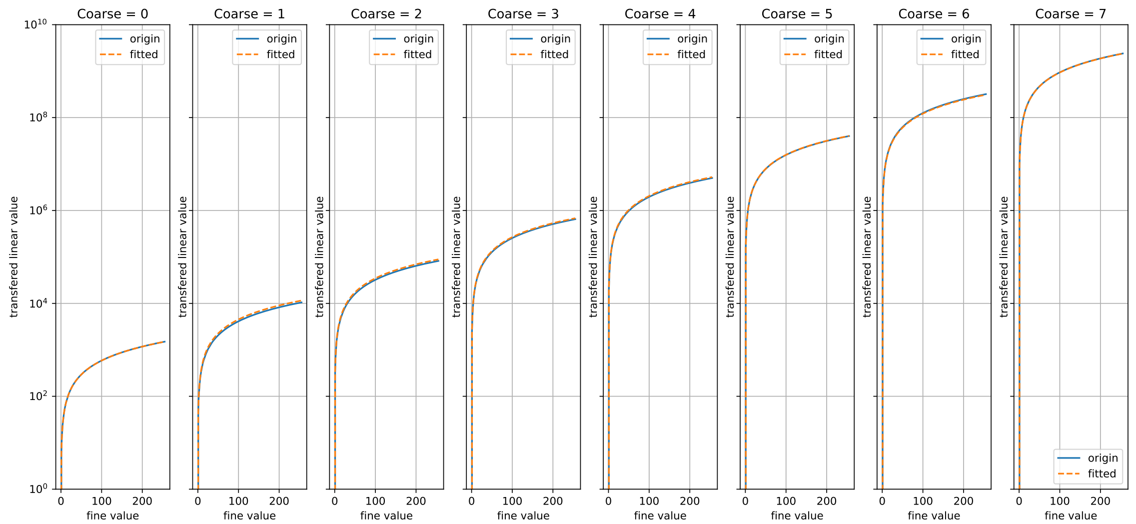
Table 2.1: The slope of bias values under each coarse set.



(a) shared y scale representation



(b) flexible y scale representation



(c) log y scale representation

Figure 2.1: cf to linear values of board 000020, fitted with function $y = 5.88e^{2.04c}f$ (c: coarse value, f: fine value)

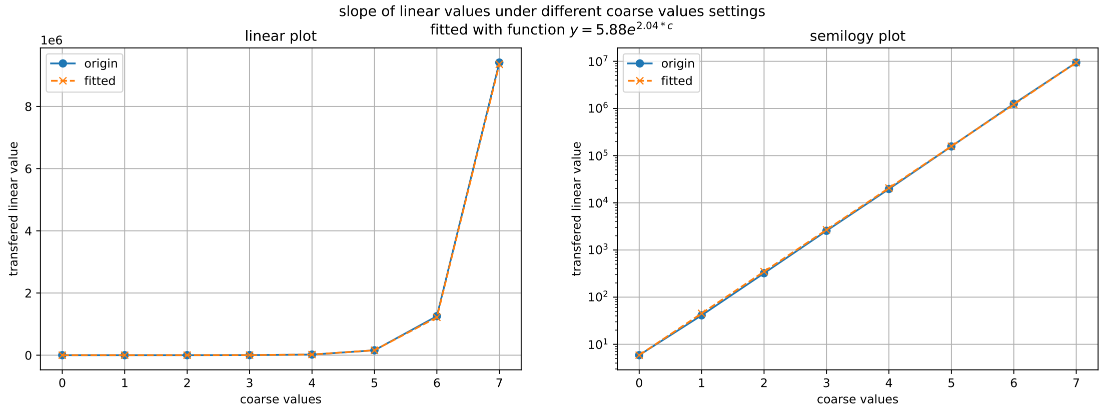
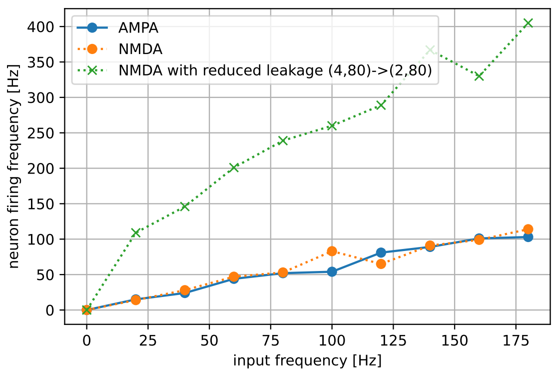
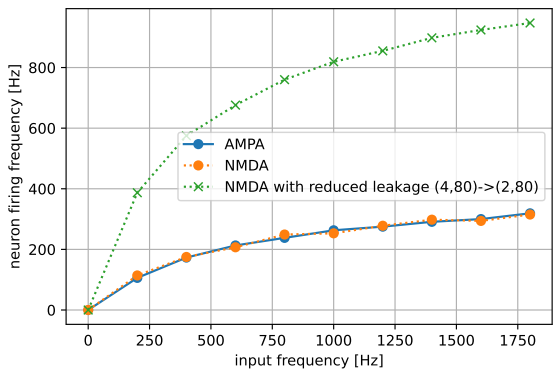


Figure 2.2: Slope of bias values under different coarse settings, fitted with $y = 5.88e^{2.04c}$



(a) Frequency range 0-180Hz



(b) Frequency range 0-1800Hz

Figure 2.3: The neuron FF curve with same AMPA, NMDA parameters applied.

NG_c	NG_f	NL_c	NL_f	NR_c	NR_f	aG_c	aG_f	aL_c
5	80	4	80	4	128	4	80	4
aL_f	aW_c	aW_f	nG_c	nG_f	nL_c	nL_f	nW_c	nW_f
80	7	100	4	80	4	80	7	100

Table 2.2: Comparison of neuron firing behaviour by giving AMPA and NMDA the same parameter.(N: Neuron, a: AMPA, n: NMDA, G: gain, L: leakage, R: refractory period, W: weight, c: coarse, f: fine)

chip	core	NG_c	NG_f	NL_c	NL_f	NR_c	NR_f
0	0	5	80	4	80	4	128
0	1	5	80	4	80	4	128
0	2	5	80	4	80	4	128
chip	core	aG_c	aG_f	aL_c	aL_f	aW_c	aW_f
0	0	4	80	4	80	4	80
0	1	4	80	4	80	6	50
0	2	4	80	4	80	5	200
chip	core	nG_c	nG_f	nL_c	nL_f	nW_c	nW_f
0	0	4	80	4	80	4	80
0	1	4	80	4	80	3	50
0	2	4	80	4	80	7	50
chip	core	AG_c	AG_f	AL_c	AL_f	AW_c	AW_f
0	0	4	80	4	80	4	80
0	1	4	80	2	200	7	200
0	2	4	80	4	80	7	20

Table 2.3: Parameter settings (N: Neuron, a: AMPA, n: NMDA, A: GABA_A, G: gain, L: leakage, R: refractory period, W: weight, c: coarse, f: fine)

simulation, the two synapse on board is considered to be the same.

To obtain a slow NMDA receptor behaviour given the same weight parameter, further reducing leakage current is required. The effect of reducing the leakage current is shown in the figure 2.3. Both reduction of leakage current and increase the weight can help increase the neuron firing rate. As shown in figure 2.3b, by increasing the input frequency range from 180 to 1800Hz, the neuron refractory period effect appears where a neuron firing rate saturation shows.

2.1.3 Parameter tuning

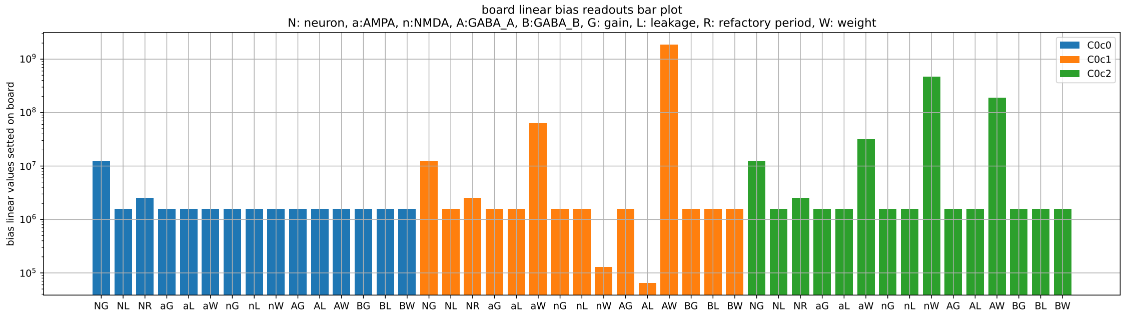
The detailed parameters applied on board to obtain a point attractor behaviour is shown in table **??**. To simplify, the major parameters tuned for core 1 is AMPA weight, NMDA weight, GABA_A weight and leakage; the major parameters tuned for core 2 is AMPA weight and GABA_A weight as visualized in figure 2.4.

To obtain a point attractor behaviour, the AMPA weight for core 1 is set as (6,50), NMDA weight is set as (3, 50), GABA_A weight is set as (7, 200) and GABA_A leakage is set as (2, 200).
The intension behind the settings is to enhance the E_{att} population recurrent connections and balance the E_{bkg} population influence. Besides, to increase the sensitivity of E population to the inhibitory afferent.

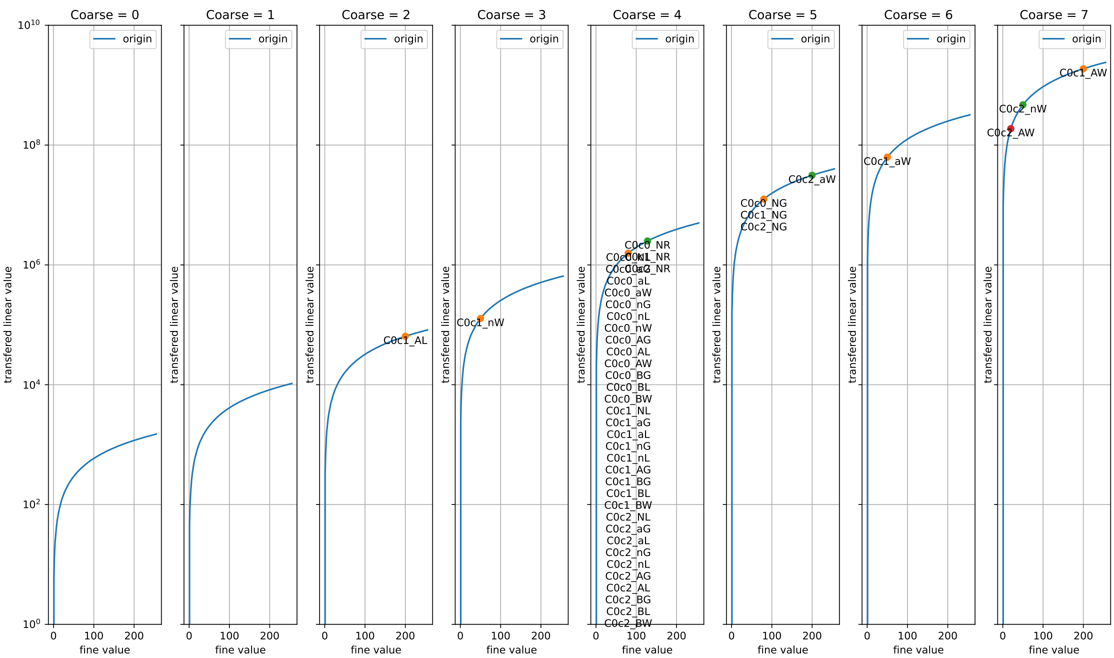
The AMPA weight for core 2 is set as (5, 200), GABA_A is set as (7, 20).
The intension behind the settings is to enhance the I population recurrent connections and increase the sensitivity to the excitatory afferent. In comparison with E population, the I population therefore obtain a faster response.

2.2 Architecture of the Point Attractor

A network structure that is similar to the point attractor network mentioned above in Chapter 1 is implemented on the DYNAPSE board as shown in figure 2.5. The DYNAP-SE1 board is



(a) Bias values bar plot



(b) Bias values comparison on curve

Figure 2.4: Applied bias values on board.(N: Neuron, a: AMPA, n: NMDA, A: GABA_A, G: gain, L: leakage, R: refractory period, W: weight, c: coarse, f: fine)

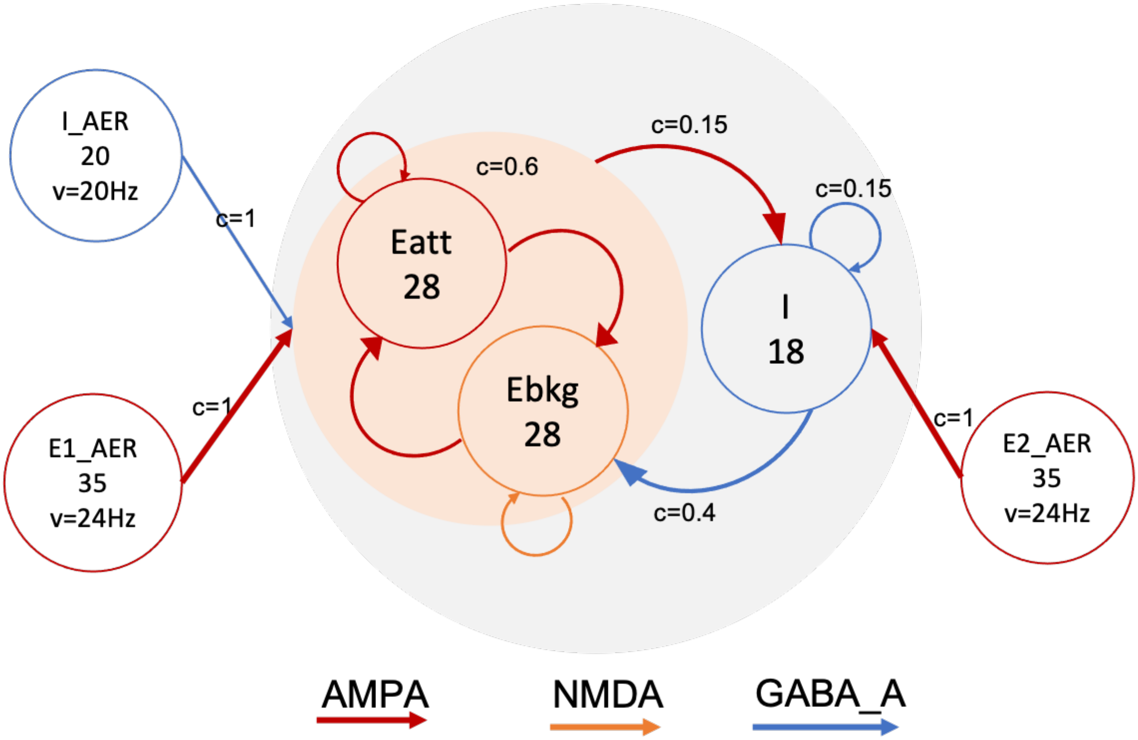


Figure 2.5: Network structure implemented on DYNAPSE board. The neurons in the gray circle is implemented on chip. The AER stimuli are from FPGA spike generator.

a neuromorphic chip which has 4 chips and each chip contains 4 cores featuring 256 silicon neurons.

By using two types of excitatory receptors (synapse) "AMPA" and "NMDA" on board, the potentiated excitatory synapse and depressed excitatory synapse are represented. Besides, the GABA_A receptor (synapse) on board is used to represent the inhibitory synapse.
The E_{att} and E_{bkg} populations connections and E and I population connection is similar to the connection mentioned in the literature [1].

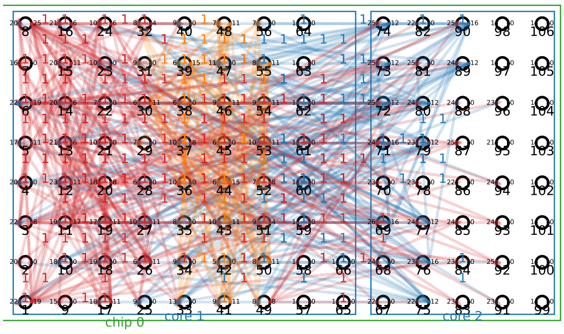
However, due to the limitation of hardware supported number of synapses that each neuron can be connected to. It is hard to build such a large network with 48 E_{att} neurons and 48 E_{bkg} neurons on this board. Therefore, a scaling of this network is done by using a factor 0.6 to ensure the connection is implementable on this chip. As a result, the number of E_{att} and E_{bkg} neurons is scaled to 28 and the number of I neurons is scaled to 18. With this scaling, the connection can be built successfully with the same connection probabilities mentioned in the literature.

Besides, instead of using big number of neurons for the input Poisson stimuli and with a very small connection probability. A small number of neuron is chosen and a probability of 1 is given for establishing the connection between the input spike generators and E, I neuron populations.

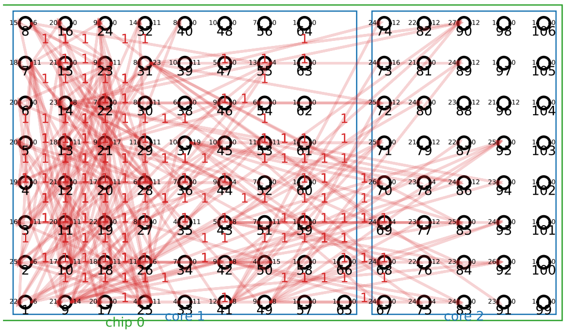
2.3 Mapping the network on the DYNAPSE board

The detailed on board connections is visualized in the figure 2.6.

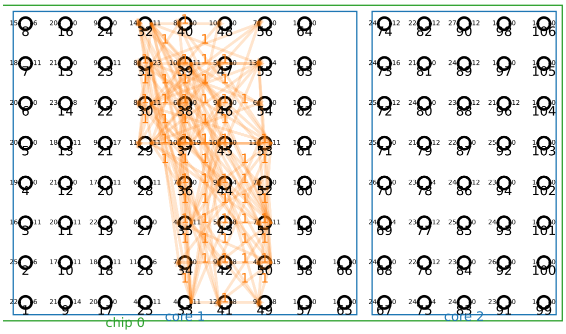
As shown in the figure, there are 106 neurons are allocated in core 0 for generating the spike input. The 56 excitatory neurons are allocated in core 1 and 18 inhibitory neurons are allocated in core2.



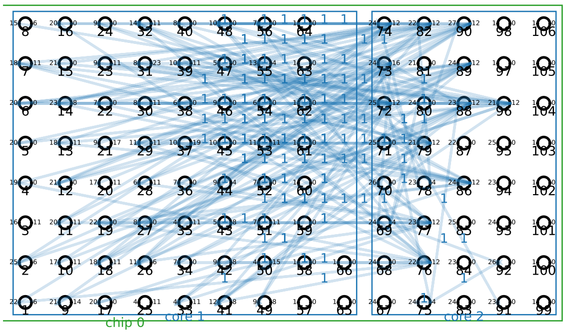
(a) All the inter-connections between E and I neuron populations.



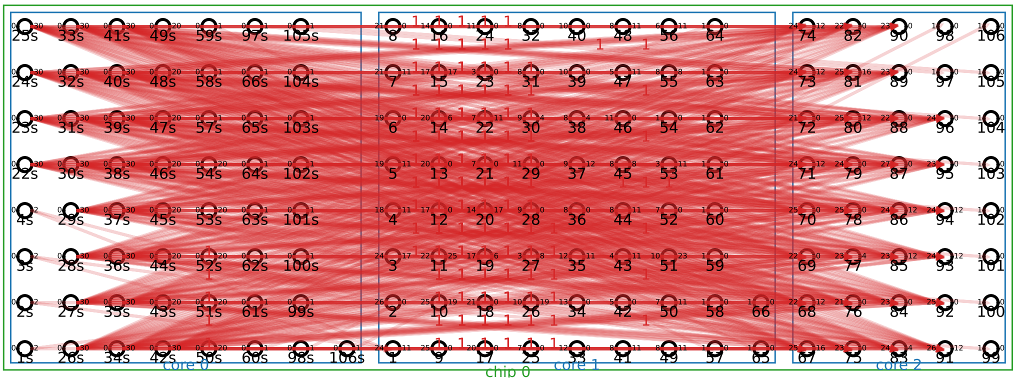
(b) All AMPA connections between E and I neuron populations.



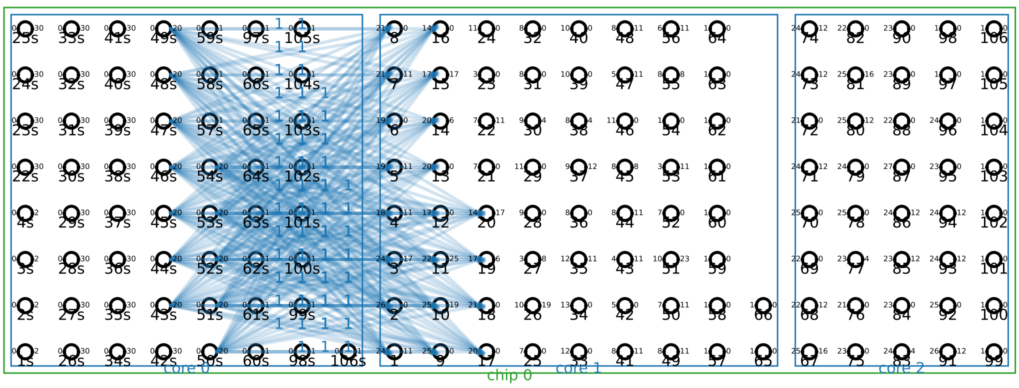
(c) All NMDA connections inside E_{bkg} neuron population.



(d) All GABA_A connections inside I neuron population and between E and I neuron population.



(e) Excitatory spike generator connected to all the E_{att} population and I neuron population.with AMPA



(f) Inhibitory spike generator connected to E_{att} neurons with GABA_A

Figure 2.6: Connections implemented on board (AMPA: red, NMDA: orange, GABA_A: blue).

There are three types of connections between the neurons that has been applied as shown in figure 2.6a, 2.6b, 2.6c and 2.6d. Different type of connection is shown with different colours where red represents AMPA, orange represents NMDA and blue represents GABA_A.
The incoming connections is shown on the left port of the neuron and the output connections is shown on the right port of the neuron. The number of fan-in of each neuron is physically limited to 64 and the numbers of input and output connections are displayed on left and right port of the neurons.

When a recurrent connection appears, the neuron will connect its left and right port, the number above the line shows the number of synapse between two nodes.

The input Poisson stimuli are generated by spike generators. Each of them are connected to each neuron in corresponding populations (E or I population). There are two types of input stimuli transmission, one is transmitted with excitatory connections AMPA as shown in figure 2.6e, and the other is transmitted with inhibitory connections NMDA as shown in figure 2.6f.

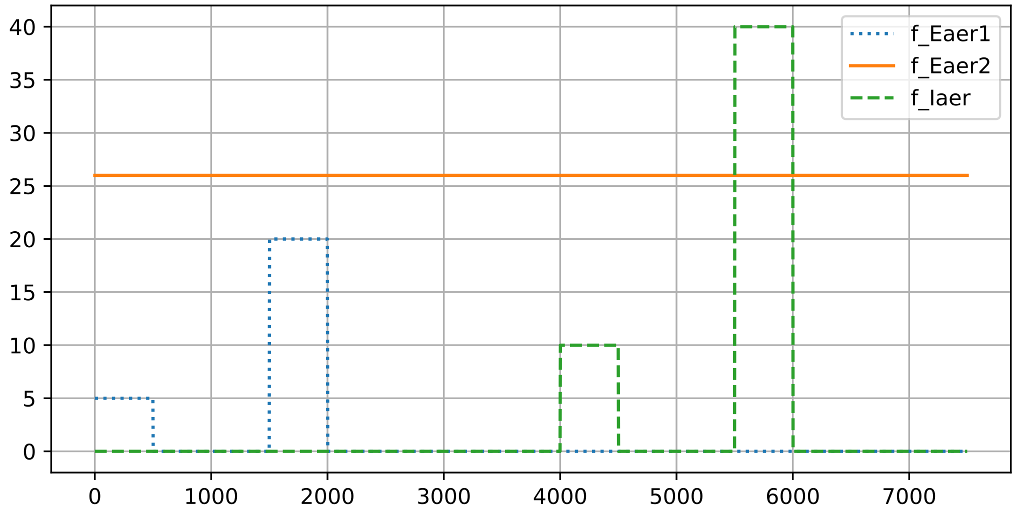
2.4 Results

By fine tuning the parameters, a final point attractor state of the E population is obtained as shown in figure 2.7. The raster plot shows three neuron populations: I, E_{bkg} and E_{att} population. To be more clear about the population firing states, the neuron population firing rate is further counted and averaged in figure 2.8.

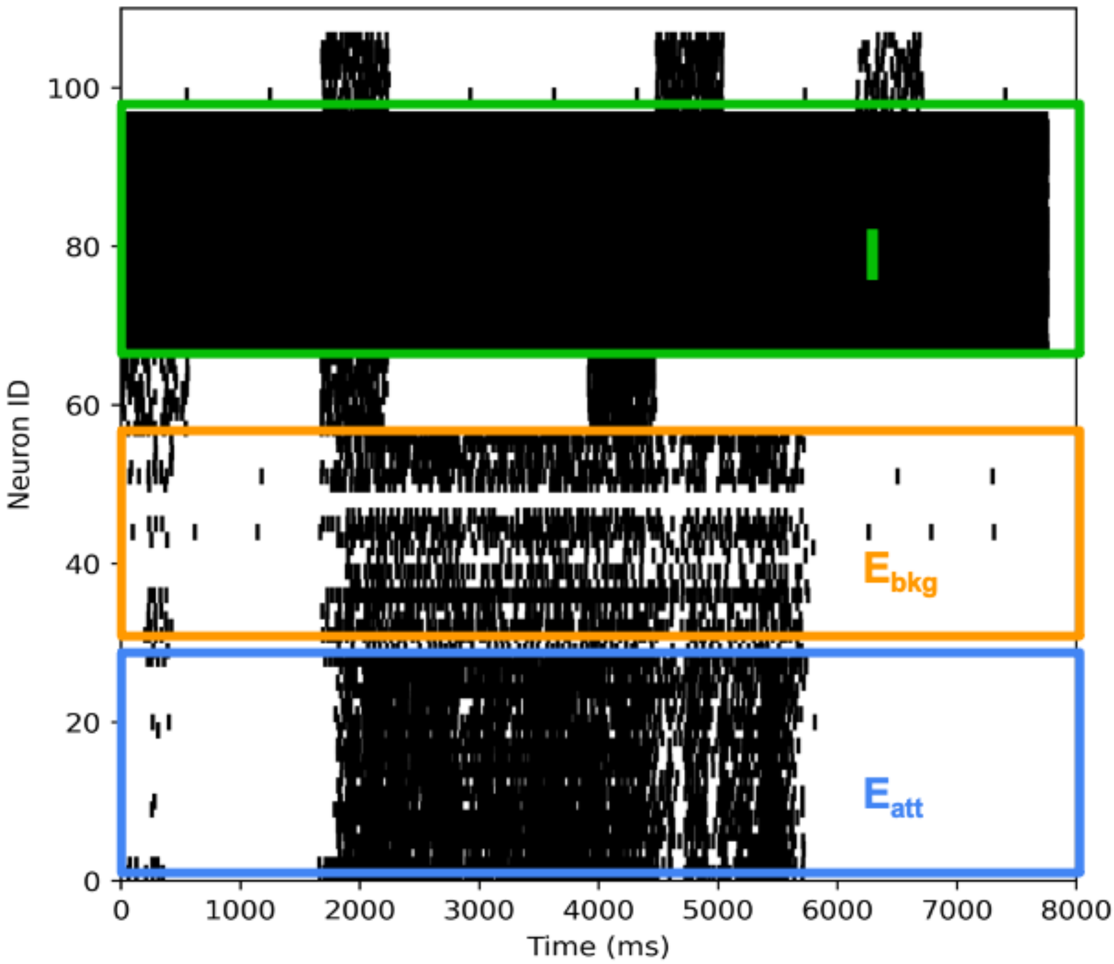
As shown in the figure,

- at time 0 ms, a weak excitation is applied to E_{att} populations, the two populations has a firing rate about 3-5 Hz.
- at time 1500 ms, a strong excitation is applied to E_{att} populations, E_{att} reaches a firing rate around 30Hz and E_{bkg} reaches a firing rate around 18 Hz.
- at time 4000 ms, a weak inhibition is applied to E_{att} populations, both E_{bkg} and E_{att} populations firing rate drop a bit but soon return back around the normal state.
- at time 5500 ms, a strong inhibition is applied to E_{att} populations, both E_{bkg} and E_{att} populations firing rate drop to zero.

It is believed that the E population neurons reaches a point attractor state where it can only switch to a "high" state with a strong excitatory stimuli, and can be stable when small inhibition or weak perturbation appears. They can only switch back to "low" state by feeding in strong inhibition stimuli.
The excitatory threshold to realise "low" to "high" state is about 5Hz and the inhibitory threshold from "high" to "low" should be larger than 10Hz. The population "high" state is around 30Hz and "low" state is 0Hz.

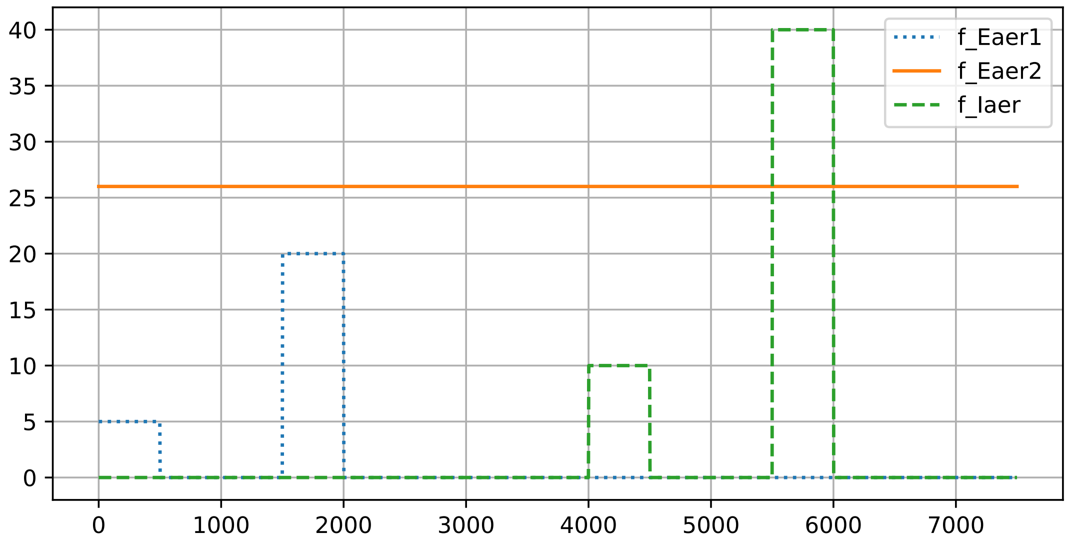


(a) Stimuli of the AER input. (Poisson spike train at different frequencies) (x: time [ms], y: frequency [Hz])

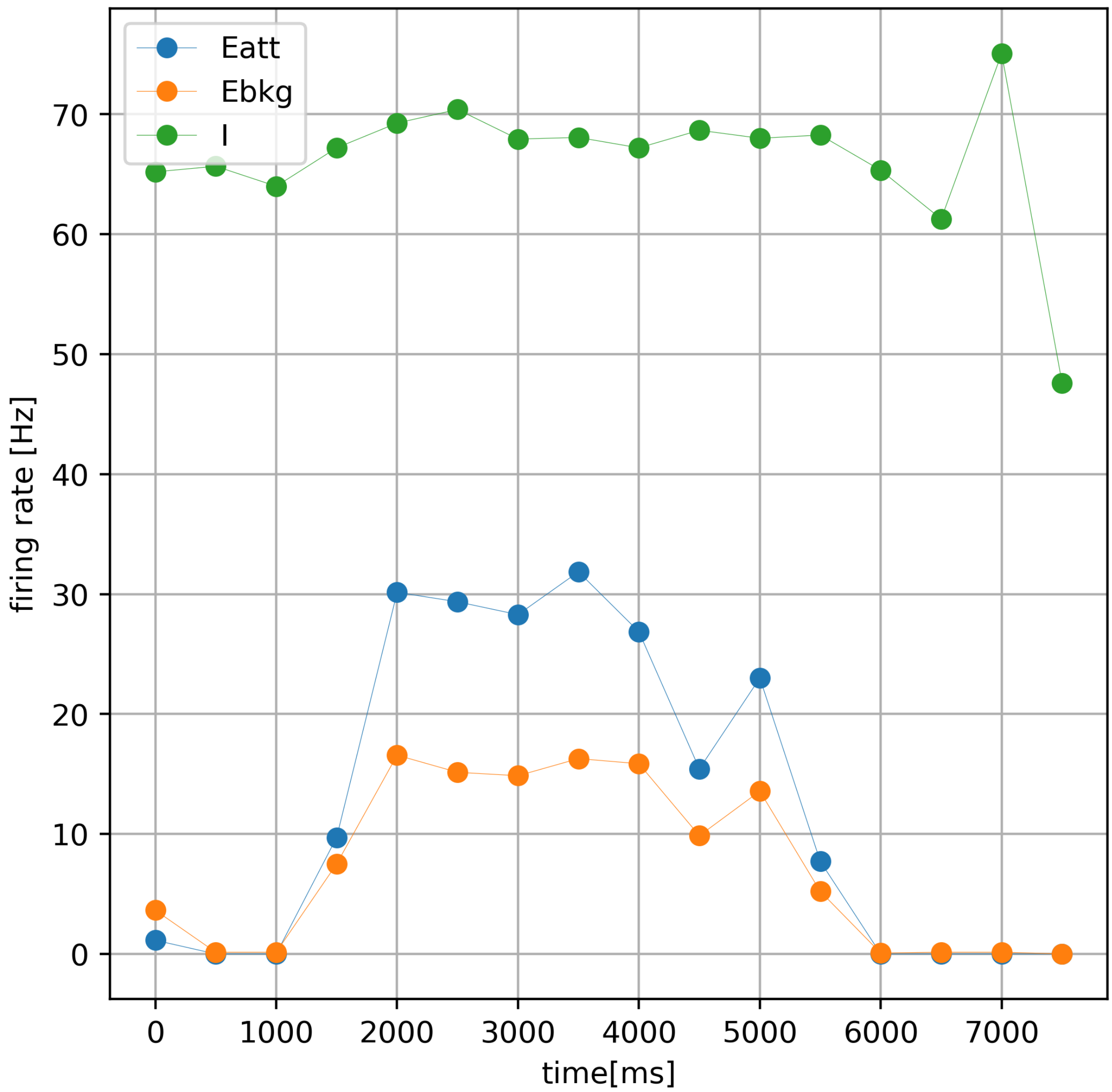


(b) (b) Raster plot of neuron activities corresponding to input stimuli in (a)

Figure 2.7: Point attractor behaviour(raster plot) of E population.



(a) Stimuli of the AER input. (Poisson spike train at different frequencies) (x: time [ms], y: frequency [Hz])



(b) Neuron population point attractor behaviour

Figure 2.8: Neuron population point attractor behaviour. At time 0ms, a weak excitation is applied to E_{att} populations, around time 1500ms, a strong excitation is applied to E_{att} population, at time 4000ms, a weak inhibition is applied to the E_{att} population, around time 5500ms, a strong inhibition is applied to the E_{att} populations.

Chapter 3

Conclusion

A point attractor network has been implemented on the DYNAPSE board by constructing a E-I balanced network and fine tuning the bias.

3.1 Network on chip

Structure

The network consists of one excitatory (E) population and one inhibitory (I) population as shown in figure 2.5. The excitatory population consists of a attractive population E_{att} and a background population E_{bkg}.
E population connect to I population with AMPA and I population propagate its signal through GABA_A. I population recurrently connect to itself through GABA_A. E_{att} and E_{bkg} population recurrently connect to themself through AMPA and NMDA respectively.
This network is a kind of derivative feedback network as shown in figure 1.1. It has been shown that derivative network is very efficient in maintaining the state of the E populations.

Size

The implemented network consists of 56 excitatory neurons and 18 inhibitory neurons (the ratio is about 3:1), 74 neurons in total compared with 127 neurons used in the literature.

Bias

To increase a neuron's firing rate, we can either increase the weight or decrease the of leakage of the synapse.
A fine tuning of the biases in different core is required to realise the balance between E population and I population. Basically, it is intended to achieve slow E population recurrent response and fast I population negative feedback[2].

Others

The relation between coarse, fine value and the linear applied bias value is measured for DYNAPSE board 020 which helps with the parameter tuning. Besides, to help validate the connections, a network connection visualisation tool is developed as shown in figure 2.6.

3.2 Point attractor behaviour

The E population will change state from "low" to "high" with relative strong stimulation and the excitatory threshold is around 5Hz. Similarly, the network is resistive to small external perturbations, the inhibitory threshold required to change the state from "high" to "low" is more than 10 Hz. The point attractor behaviour is believed to be one of the key behaviours of the working memory.

3.3 Discussions

A lot more interesting topics can be further investigated based on this implementation method, like checking the relation between network size and the attractor state, how the input stimuli duration and strength can influence the attractor state, how to concatenate several EI balanced population to realise the propagation of information in multiple layers.

Concerning the network structures, some further implementations can be made to compare the derivative feedback network with pure positive feedback structures.

Bibliography

[1] Paolo Del Giudice. Robust working memory in an asynchronously spiking neural network realized with neuromorphic VLSI. page 16.

[2] Sukbin Lim and Mark S Goldman. Balanced cortical microcircuitry for maintaining information in working memory. 16(9):1306–1314.