University of Tübingen
Graduate Training Center of Neuroscience
Neural Information Processing

Memory Consolidation and Synaptic Plasticity on the Neuromorphic BrainScaleS-2 System

Laboratory Report

presented by

Amani Atoui

The study was supervised by
Dr. Johannes Schemmel
Dr. Sebastian Billaudelle
Jakob Kaiser

Electronic Vision Group
Kirchhoff's Institute of Physics
University of Heidelberg

Duration of the lab rotation: 11 weeks
Deadline of submission: March 2024

Abstract

BrainScaleS-2 is a spiking neuromorphic platform that promises accelerated emulation of biological neural networks. As a first step towards emulating memory consolidation in recurrent neural networks that incorporate the synaptic tagging and capture (STC) hypothesis and calcium dynamics in their synaptic model, we investigate the behavior of a single synapse. We simulate the standard plasticity protocols for a single synapse that induce different plasticity mechanisms. We then account for the computational time required by the digital circuits of BrainScaleS-2 to update the synaptic weights. By simulating the model using different timescales, we show that the root-mean-squared error (RMSE) in synaptic weights obtained at a slow timescale compared to synaptic weights obtained at the original timescale can be tolerated up to a certain time beyond which the behavior of the protocols diverges from their expected behavior. We argue that the final assessment of the weight updates depends on the full-network simulation where we use memory performance measures that depend on the synaptic weights.

Contents

List of Figures

List of Tables

Acronyms

AdEx adaptive exponential integrate-and-fire.

CADC column-parallel analog-to-digital converter.

CMOS complementary metal oxide semiconductor.

IQR interquartile range.

LTD long-term depression.

LTP long-term potentiation.

PPU plasticity processing unit.

RMSE root-mean-squared error.

SIMD single instruction, multiple data.

SLFS strong low-frequency stimulation.

STC synaptic tagging and capture.

STET strong tetanic stimulation.

WLFS weak low-frequency stimulation.

WTET weak tetanic stimulation.

Introduction

With the rising need for parallel and energy-efficient computing, the standard Von Neumann computer architecture fails to satisfy these requirements due to its memory constraints and use of sequential commanding that hinder the computing performance (Schuman et al., 2022). Inspired by the human brain's architecture and connectivity, neuromorphic computing promises to fulfill these needs using a spiking neural network architecture. Neuromorphic architectures possess processing-memory collocation similar to the neuron-synapse functionality in the brain assuming the neurons and synapses to be the primary computational units (Schuman et al., 2022). Neuromorphic computing has gained even more popularity with the advancement of artificial intelligence and the demand for powerful hardware with low-power features by which it can offer significant advantages to many fields such as autonomous systems including sensors, brain machine interfaces, and robotics (Indiveri and Liu, 2015). This computing paradigm is also used to solve machine learning tasks using large-scale neural modelling or very-large-scale spiking neural networks (Indiveri and Liu, 2015). For research purposes in computational neuroscience, neuromorphic computing can speed-up the simulation of biological neural networks (Indiveri and Liu, 2015).

The neurological function we are considering in this work is memory consolidation in recurrent neural networks. Memory is a cognitive function of the brain that is associated with learning; whether gaining new knowledge or modifying existing knowledge, this knowledge is stored in the brain in the form of a memory trace (Nadel and Land, 2000). One of the established theoretical explanations to the learning-memory association is referred to as Hebbian learning which depends on the spiking activity of the presynaptic and postsynaptic neurons (Luboeinski and Tetzlaff, 2021). A consistent firing of the presynaptic neuron elicits postsynaptic spikes and drives the postsynaptic calcium concentration. Depending on this concentration, the synapse is either strengthened in a process referred to as long-term potentiation (LTP) or weakened in a process referred to as long-term depression (LTD) (Luboeinski and Tetzlaff, 2021).

Memory involves a variety of dynamics, starting with encoding which is the initial learning of information. Information is then stored by a process referred to as consolidation to maintain information over time. Memory consolidation is performed in two stages; the first stage is synaptic consolidation, which involves local molecular processes and morphological changes of the synapse (Lamprecht and LeDoux, 2004). The second stage is systems consolidation which occurs at a higher level, mainly between the hippocampus and neocortex to maintain information for a prolonged period (Luboeinski and Tetzlaff, 2021). Finally, the stored information is accessed through retrieving or recalling (Luboeinski and Tetzlaff, 2021).

More details on plasticity are explained in the theory on synaptic consolidation, in which both plasticity mechanisms, LTP and LTD, involve two phases (Lamprecht

and LeDoux, 2004). The early phase is characterized by the increase in calcium concentration, while the late phase is characterized by insertion of neurotransmitter receptors in the case of LTP and their removal in the case of LTD (Lamprecht and LeDoux, 2004). However, the transfer from the early phase to the late phase was not resolved for a long time. Plasticity is input-specific and occurs at specific dendrites of the neuron, while proteins necessary for late-phase plasticity are synthesized and trafficked from inside the cell to this specific synapse, so it was unclear how the neuron can identify the synapse that should be strengthened or weakened (Luboeinski and Tetzlaff, 2021). To explain this transfer, the synaptic tagging and capture (STC) hypothesis was proposed. This hypothesis states that early-phase plasticity initiates the creation of a short-lasting protein-synthesis-independent 'synaptic tag' as a potential for a long-lasting change in synaptic efficacy but not a commitment itself (Frey and Morris, 1997). If the proteins needed for the late-phase transfer are sufficient, the tagged synapse captures these proteins and late-phase plasticity is induced (Frey and Morris, 1997). Consequently, this tag serves as the indicator for the neuron to know which synapse is involved in the plasticity mechanism, but the tag does not guarantee the transfer to the late phase.

The STC hypothesis was able to explain memory consolidation in single synapses and feedforward neural networks, but only recently Luboeinski and Tetzlaff (2021) verified that STC can explain and improve memory consolidation in recurrent neural networks. Recurrent networks are the most prevalent networks in the brain, especially when it comes to memory functions, so understanding the underlying mechanisms behind recurrent-network models is essential for understanding memory functions in the brain.

In this work, we aim to emulate memory consolidation in biological networks on the neuromorphic BrainScaleS-2 system using the STC hypothesis and calcium dynamics for plasticity. We rely on the simulations and approach provided by Luboeinski and Tetzlaff (2021) in their work on memory consolidation in recurrent neural networks. We focus on single-synapse behavior and start by reproducing simulations of single-synapse plasticity using standard plasticity experiments. We then define the constraints imposed by the emulation of a single synapse on BrainScaleS-2 (Pehle et al., 2022). Specifically, we account for the computation time required by the digital circuits to update synaptic weights by reproducing the simulations at slower time steps. To assess the feasibility of this approach, we use the root-mean-squared error (RMSE) to compare the synaptic weights obtained at the slow time steps with the synaptic weights obtained using the simulation time step defined in (Luboeinski and Tetzlaff, 2021). We conclude that using slower timescales is a feasible solution for the hardware emulation but should be further assessed during the full-network

simulation.

Methods

2.1 Single-Synapse Simulation

2.1.1 Plasticity Scheme

In this section, we describe the neurological mechanisms behind the used synaptic model in fig. 2.1. The model incorporates the calcium dynamics according to the STC hypothesis. In this model, spikes arriving from presynaptic neuron j and occurring at times t_j^n along with the total synaptic weight w_{ji} affect the postsynaptic membrane potential v_i of postsynaptic neuron i. If the membrane potential exceeds a threshold, a postsynaptic spike is elicited at time t_i^m. The presynaptic and postsynaptic spikes drive the postsynaptic calcium concentration c_{ji} which induces early-phase plasticity, quantified by the early-phase weight h_{ji}. If the synapse is tagged, and the protein amount synthesized p_i during the early phase is sufficient, late-phase plasticity is induced, quantified by the late-phase weight z_{ji}.

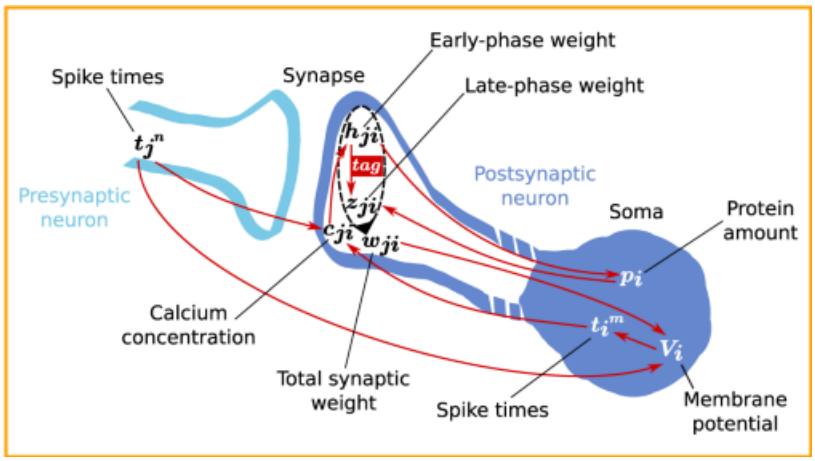

Figure 2.1: The synaptic model integrating various mechanisms of calcium dependent synaptic plasticity and the STC hypothesis. Modified from Luboeinski and Tetzlaff (2021).

2.1.2 Model Equations

All equations governing the single-synapse plasticity are adopted from Luboeinski and Tetzlaff (2021). The dynamics of the postsynaptic membrane potential follow

the integrate-and-fire neuron model and are described in eq. (2.1):

$$\tau_{\mathrm{mem}} \frac{dV_i(t)}{dt} = V_{\mathrm{rev}} - V_i(t) + \mathrm{R} \cdot (\mathrm{I_{bg}(t)} + \mathrm{I_{stim}(t)} + \mathrm{I_{syn,i}(t)}) \tag{2.1}$$

with reversal potential V_{rev}, membrane time constant τ_{mem}, membrane resistance R, external background current $\mathrm{I_{bg}(t)}$, external stimulus current $\mathrm{I_{stim}(t)}$, and synaptic current $\mathrm{I_{syn,i}(t)}$. The synaptic current $\mathrm{I_{syn}}$ is defined as:

$$\mathrm{I_{syn,i}(t)} = \sum_j \sum_{t_j^k} w_{ji} \cdot \exp\left(-(\mathrm{t} - \mathrm{t_j^k} - \mathrm{t_{ax,delay}})/\tau_{\mathrm{syn}}\right) \tag{2.2}$$

with axonal time delay $\mathrm{t_{ax,delay}}$ and synaptic time constant τ_{syn}. For a single synapse, the background current and the stimulation current are set to zero. The membrane potential for the single synapse thus reduces to eq. (2.3).

$$\tau_{\mathrm{mem}} \frac{dV_i(t)}{dt} = V_{\mathrm{rev}} - V_i(t) + \mathrm{R} \cdot \mathrm{I_{syn,i}(t)} \tag{2.3}$$

The calcium dynamics follow eq. (2.4) with τ_{c} being the calcium time constant, c_{pre} being the contribution of presynaptic spikes, c_{post} being the contribution of postsynaptic spikes, and $t_{\mathrm{c,delay}}$ being the delay of calcium concentration triggered by presynaptic spikes.

$$\frac{dc_{ji}(t)}{dt} = -\frac{c_{ji}(t)}{\tau_{\mathrm{c}}} + c_{\mathrm{pre}} \sum_n \delta(t - t_j^n - t_{\mathrm{c,delay}}) + c_{\mathrm{post}} \sum_m \delta(t - t_i^m), \tag{2.4}$$

The dynamics of the early-phase weight are governed by eq. (2.5) with $\Theta[.]$ being the Heaviside function and τ_h being a time constant. The first term of eq. (2.5) describes a relaxation of the early-phase weight to its initial value h_0, the second term describes early-phase LTP with rate γ_{p} for calcium concentration above the threshold θ_{p}, and the third term describes early-phase LTD with rate γ_{d} for calcium concentration above the threshold θ_{d}.

$$\begin{aligned} \tau_h \frac{dh_{ji}(t)}{dt} = {} & 0.1\ (h_0 - h_{ji}(t)) + \gamma_{\mathrm{p}}(1\ \mathrm{nC} - h_{ji}(t)) \cdot \Theta[c_{ji}(t) - \theta_{\mathrm{p}}] \\ & - \gamma_{\mathrm{d}} h_{ji}(t) \cdot \Theta[c_{ji}(t) - \theta_{\mathrm{d}}] + \xi(t), \end{aligned} \tag{2.5}$$

The term $\xi(t) = \sqrt{\tau_h[\Theta(c_{ji}(t) - \theta_{\mathrm{p}}) + \Theta(c_{ji}(t) - \theta_{\mathrm{d}})]}\ \sigma_{\mathrm{pl}}\ \Gamma(t)$ describes the calcium-dependent noise-driven fluctuations with standard deviation σ_{pl}, and Gaussian white noise $\Gamma(t)$ with mean zero and variance $\frac{1}{\Delta t}$, where Δt is the time step for numerical computations. Knowing the early-phase weight, the protein amount is updated using eq. (2.6) where α is the protein synthesis rate and θ_{pro} is the protein synthesis threshold.

$$\tau_p \frac{dp_i(t)}{dt} = -p_i(t) + \alpha\ \Theta\left[\left(\sum_j |h_{ji}(t) - h_0|\right) - \theta_{\mathrm{pro}}\right] \tag{2.6}$$

The dynamics of the late-phase weight depend on the protein amount, early-phase weight, and a tagging threshold θ_{tag}.

$$\begin{aligned} \tau_z \frac{dz_{ji}(t)}{dt} = {} & p_i(t) \cdot (1 - z_{ji}(t)) \cdot \Theta[(h_{ji}(t) - h_0) - \theta_{\mathrm{tag}}] \\ & - p_i(t) \cdot (z_{ji} + 0.5) \cdot \Theta[(h_0 - h_{ji}(t)) - \theta_{\mathrm{tag}}] \end{aligned} \tag{2.7}$$

Finally, the total synaptic weight is given by:

$$w_{ji}(t) = h_{ji}(t) + h_0 \cdot z_{ji}(t) \tag{2.8}$$

2.1.3 Plasticity Protocols

Stimulating the synapse is performed through Poisson presynaptic spikes. To account for different plasticity forms of the synapse, we use the standard plasticity protocols which vary by strength and frequency described in fig. 2.2 for stimulating a single synapse. Tetanic (high-frequency) stimulations induce potentiation due to the high calcium concentration while low-frequency stimulations induce depression due to the moderate calcium concentration. On the other hand, strong stimulations cause the tagging of the synapse and the synthesis of a sufficient protein amount whereas weak stimulations only cause the synapse tagging without a sufficient synthesis of protein amount. Consequently, the strong tetanic stimulation (STET) protocol induces early-phase and late-phase LTP whereas weak tetanic stimulation (WTET) protocol induces only early-phase LTP. On the other hand, strong low-frequency stimulation (SLFS) protocol induces early-phase and late-phase LTD whereas weak low-frequency stimulation (WLFS) protocol induces only early-phase LTD.

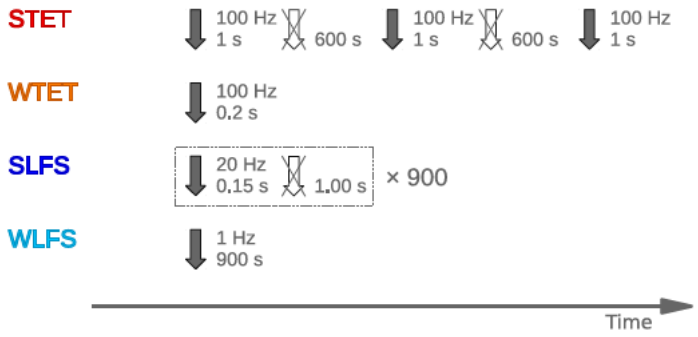

Figure 2.2: Standard protocols for the the induction of early- and late-phase synaptic potentiation and depression. From supplementary information of Luboeinski and Tetzlaff (2021).

2.1.4 Simulation Scheme and Parameters

For the differential equations defined in section 2.1.2, we use the explicit Euler method to update all the model parameters (Kong et al., 2021). More specifically, let $\frac{dy(t)}{dt} = F(t, y)$ be a first-order differential equation. The linear approximation of $y(t)$ at t_{n+1} around t_n is:

$$y(t_{n+1}) = y(t_n) + (t_{n+1} - t_n) \cdot \left.\frac{dy}{dt}\right|_{t=t_n} \tag{2.9}$$

For a regular time step $\Delta t = t_{n+1} - t_n$, we can re-write the explicit Euler formula as (Kong et al., 2021):

$$y(t_n + \Delta t) = y(t_n) + \Delta t \cdot F(t_n, y(t_n)) \tag{2.10}$$

By following the plasticity scheme in section 2.1.1, the weights for the different protocols are computed for a total duration of 8 hours and a time step $\Delta t = 0.2$ ms. Poisson presynaptic spikes are simulated according to the plasticity protocols in section section 2.1.3, and the parameters presented in the equation are provided in table 2.1 (Luboeinski and Tetzlaff, 2021).

Symbol	Value	Description
Δt	0.2 ms	Duration of one time step for numerical computation
τ_{mem}	10 ms	Membrane time constant
τ_{syn}	5 ms	Synaptic time constant
$t_{ax,delay}$	3 ms	Axonal spike delay
t_{ref}	2 ms	Refractory period
R	10 MΩ	Membrane resistance
V_{rev}	−65 mV	Reversal potential
V_{reset}	−70 mV	Reset potential
V_{th}	−55 mV	Threshold potential for spiking
h_0	0.420 075 nC	Initial early-phase weight
$t_{c,delay}$	0.0188 s	Delay of postsynaptic calcium influx after presynaptic spike
c_{pre}	1	Presynaptic calcium contribution
c_{post}	0.2758	Postsynaptic calcium contribution
τ_c	0.0488 s	Calcium time constant
τ_p	60 min	Protein time constant
τ_z	60 min	Late-phase time constant
γ_p	1645.6	Potentiation rate
γ_d	313.1	Depression rate
θ_p	3	Calcium threshold for potentiation
θ_d	1.2	Calcium threshold for depression
σ_{pl}	0.290 436 nC $\mathrm{s}^{-1/2}$	Standard deviation for plasticity fluctuations
α	1	Protein synthesis rate
θ_{pro}	0.210 037 nC	Protein synthesis threshold
θ_{tag}	0.084 014 9 nC	Tagging threshold

Table 2.1: Neuron and synapse model parameters. Adapted from Luboeinski and Tetzlaff (2021).

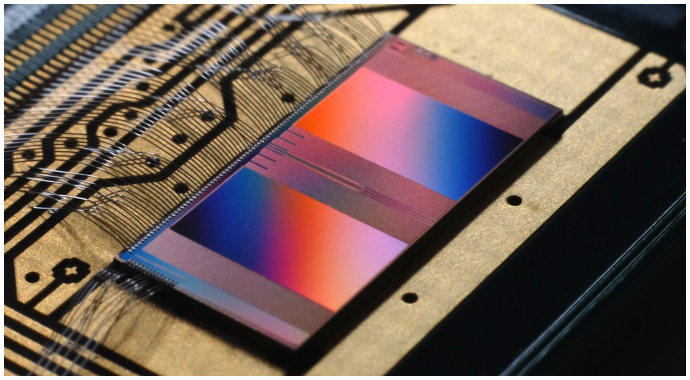

Figure 2.3: Photograph of the full-size BrainScaleS-2 chip. Modified from Billaudelle (2022).

2.2 Hardware Emulation

2.2.1 BrainScaleS-2

BrainScaleS-2 (fig. 2.3) is a neuromorphic chip enclosing 512 silicon neurons where each neuron can form up to 256 synapses (Pehle et al., 2022) and has dynamics that follow the adaptive exponential integrate-and-fire (AdEx) model proposed by Brette and Gerstner (2005). BrainScaleS-2 possesses several features that allow fast emulation of biological neural networks. First and foremost, the neuron dynamics evolve at a 1000-fold accelerated timescale relative to real time due to the time constants of the complementary metal oxide semiconductor (CMOS) technology it uses (Pehle et al., 2022). Second, the chip adapts a hybrid plasticity scheme which combines an analog circuit for emulating neuron and synapse dynamics and a digital circuit for control and calculations (Pehle et al., 2022). Applying these concepts for memory consolidation, the stimulating presynaptic spikes and the resulting postsynaptic spikes will be injected to the analog circuit which will emulate the neuronal dynamics, and the digital circuit will calculate the synaptic weights and protein amount. Concerning the calcium concentration, the link between the spike occurrences and the weight updates, it will be computed using a special feature of the analog circuit, but its details will not be discussed in this report.

2.2.2 Constraints

The analog circuit requires the total synaptic weight to emulate the neuronal dynamics shown in eq. (2.1) for the postsynaptic membrane potential. Since this weight is updated regularly by the plasticity rule, the postsynaptic spikes should be read-out and the calcium concentration should be sampled to compute the protein amount and update the synaptic weights. Specifically, the calcium concentration will be computed by a special feature of the analog circuit, so it is an analog signal that needs to be converted to a digital signal using the column-parallel analog-to-

digital converter (CADC). It has been established in benchmarking procedures for BrainScaleS-2 that a single CADC read-out requires 1.7 µs assuming the default plasticity processing unit (PPU) clock frequency of 250 MHz (Billaudelle, 2022). The CADC performs the read-out in parallel for up to 256 neurons (Billaudelle, 2022).

For computing the synaptic weights and protein amount, the single instruction, multiple data (SIMD) unit can calculate 128 8-bit or 64 16-bit integers in parallel (Pehle et al., 2022). We need to update four parameters per neuron, namely the early-phase weight, protein amount, late-phase weight, and total synaptic weight. The operations performed by these updates using the Euler formula (eq. (2.10)) require 50 ns per instruction. After computing the weights and protein amount, these need to be updated through write instructions performed by the PPU. Each write instruction can update weights for 256 synapses per single neuron in 160 ns, so this has to be repeated for 256 neurons by each PPU. An estimate of the time consumed by the operations required for weight updates is presented in table 2.2.

Operation	Involved Hardware Unit	Number of Instructions per Operation	Time per Instruction	Total Time
Analog read-out and sampling	CADC	1	1.7 µs	1.7 µs
Early-phase weight computation	SIMD	17	50 ns	850 ns
Late-phase weight computation	SIMD	16	50 ns	800 ns
Protein amount computation	SIMD	9	50 ns	450 ns
Total synaptic-weight computation	SIMD	2	50 ns	100 ns
Writing weights	PPU	256	160 ns	40.96 µs

Table 2.2: Required time by BrainScaleS-2 to perform operations necessary for synaptic weight updates at 250 MHz clock frequency. An operation involves 256 neurons.

The time step used in the simulation was $\Delta t = 0.2$ ms which translates to 0.2 µs on BrainScaleS-2 if the dynamics are maintained. The organization of instructions and pipelining should be further investigated to have a clear estimate of the required time. Furthermore, the network is sparsely coupled, and the probability of connectivity is 0.1 (Luboeinski and Tetzlaff, 2021), which might reduce the time required for computations and writing weights. Given these two considerations, more information and benchmarking are needed to exactly estimate the total time required for weight updates by BrainScaleS-2 for the full network emulation. Nevertheless, the information in table 2.2 clearly shows that 0.2 µs is not sufficient to perform

read-outs, sampling, calculations, and writing.

To resolve this issue, we propose using a slower timescale for updating the synaptic weights on BrainScaleS-2. We define a new simulation scheme to study the effect of the slow update on the synaptic weights (fig. 2.4). In this simulation, the membrane potential and calcium concentration are updated at the base timescale ($\Delta t = 0.2$ ms) since they will be emulated by the analog circuit of BrainScaleS-2. The protein amount and synaptic weights are updated at a slow timescale since they will be computed by the digital circuits of BrainScaleS-2. We refer to the timescale at which we update the protein amount and synaptic weights as an "update time". The simulation is repeated for 100 trials for 5 different update times ranging from 10 ms to 300 ms.

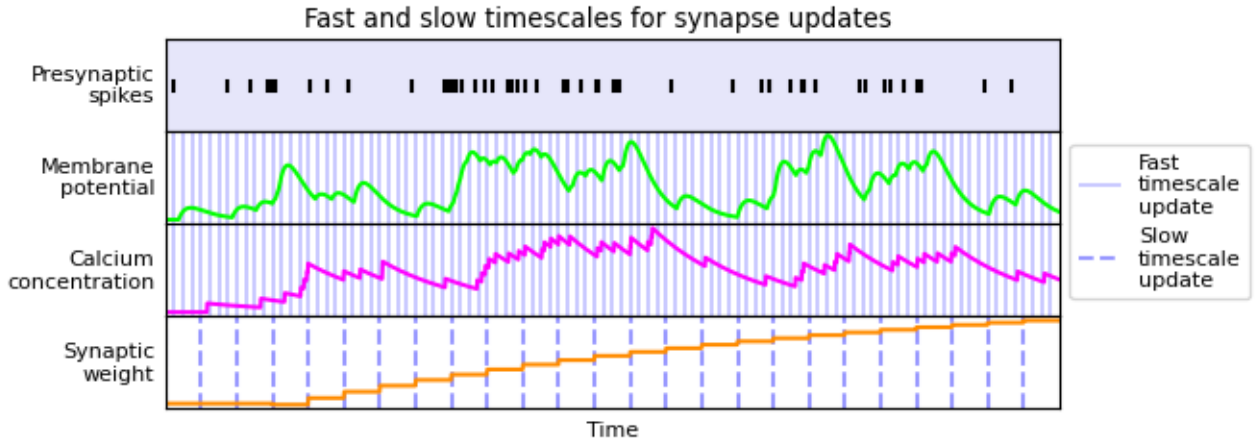

Figure 2.4: Fast and slow timescales for single-synapse simulation. Presynaptic spikes correspond to continuous-time dynamics that will be injected to the analog core of BrainScaleS-2. The membrane potential and calcium concentration are updated at a fast timescale in the simulation since they will be emulated on the analog core of BrainScaleS-2. Synaptic weights will be computed by the digital circuits of BrainScaleS-2. These weights have long time constants, so it is expected that updating them at slow timescales would still achieve the target behavior.

2.2.3 Measures of Performance

The RMSE at each update time is used to quantify the error in synaptic weights and protein amount that results from using slower time updates. The ground truth for calculating the RMSE are the values of synaptic weights and protein amount obtained at the base timescale ($\Delta t = 0.2$ ms) for the same random seed. Assessing the total synaptic weight is also useful since it plays a significant role in the behavior of the full network. According to eq. (2.7), the total synaptic weight couples the early-phase weight with the late-phase weight. The early-phase weight always converges to the same equilibrium value, so it is not as informative to assess its change with time. It is more convenient to look at the final value of the late-phase weight

as it is decoupled from the early-phase weight.

Results

3.1 Single-Synapse Simulation using the Base Timescale

The results of reproducing the single-synapse simulation for the four different plasticity protocols are shown in fig. 3.1. Each protocol was executed 100 times to generate different spike timings. The synaptic weights in fig. 3.1 agree with those obtained in the single-synapse simulations of Luboeinski and Tetzlaff (2021) number-wise and behavior-wise. All protocols induce early-phase plasticity indicated by the dynamics of the early-phase weight. On the other hand, only strong protocols induce late-phase plasticity. STET induces late-phase LTP shown by the increase in the late-phase weight while SLFS induces late-phase LTD shown by the decrease in the late-phase weight. The variability of weights due to the variability in spike timings across different trials is shown by the margins around the plots. This variability is higher in the WTET and SLFS protocols compared to STET and WLFS protocols.

3.2 Impact of Different Timescales

A boxplot of the RMSE for all 100 trials is shown in fig. 3.2. For the STET, WTET, and WLFS protocols, the RMSE for the synaptic weights and protein amount remains almost unchanged up to an update-time of 0.02 s. For the weak protocols, the late-phase weight remains zero showing a correct behavior for the protocols at slower timescales. The SLFS protocol has high RMSE values relative to the rest of the protocols even at low update times. Having a look at the final late-phase weight for the strong protocols (fig. 3.3), the median final late-phase weight for the STET protocol remains close to that at the base timescale for update times up to 0.05 s. The interquartile range (IQR) for the STET protocol is approximately 0.02 nC at the base timescale, and it remains close to this value for update times up to 0.02 s. For the SLFS protocol, the median final late-phase weight also remains close to that at the base timescale for update times up to 0.05 s. However, the IQR for the SLFS protocol is approximately 0.5 nC at the base timescale, which is higher than that of the STET protocol, but it also remains almost unchanged for update times up to 0.02 s. For update times greater than 0.05 s, the behavior of SLFS is off as the final late-phase weight reaches 0 or even positive values for some spike timings indicating the absence of late-phase plasticity or late-phase LTP.

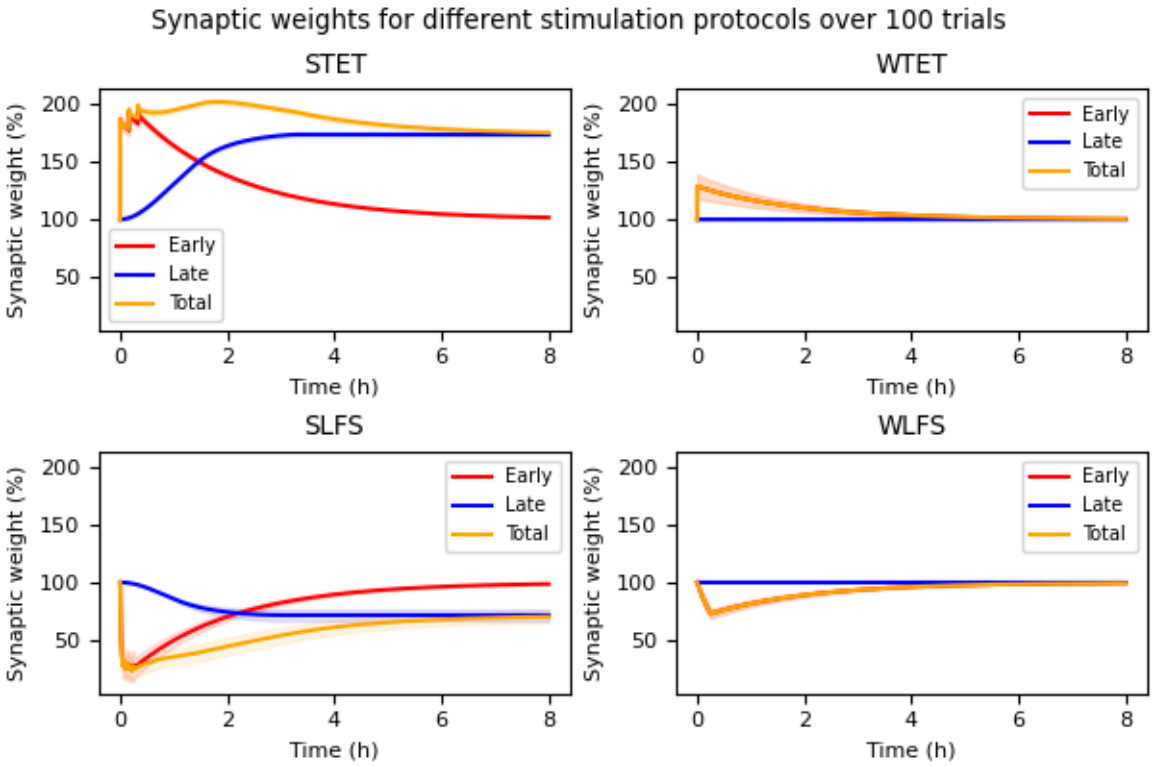

Figure 3.1: Impact of strong and weak tetanic and low-frequency stimulation protocols described in fig. 2.2 on a single synapse. The simulation is carried out for 100 trials using the synapse and neuron parameters listed in table 2.1 and a simulation time step of 0.2 ms. The lines correspond to the average weights across the 100 trials, and the bands correspond to one standard deviation from the average. The results are in agreement with those obtained in figure 2 of Luboeinski and Tetzlaff (2021)

Discussion

4.1 Simulation Results

From the single-synapse simulations we carried out, we aim towards emulating a single-synapse on the neuromorphic BrainScaleS-2 system. We first reproduced the results of the single-synapse behavior to standard plasticity protocols using a model that incorporates calcium dynamics proposed by Luboeinski and Tetzlaff (2021). The base time step of 0.2 ms used in Luboeinski and Tetzlaff (2021) is not sufficient for updating synaptic weights on BrainScaleS-2. We then accounted for this constraint in a simulation that uses fast and slow timescales, and we assessed this simulation scheme using RMSE of synaptic weights at each update time. We found out that the RMSE can be tolerated up to certain update times, but further assessment is needed with the full-network simulation.

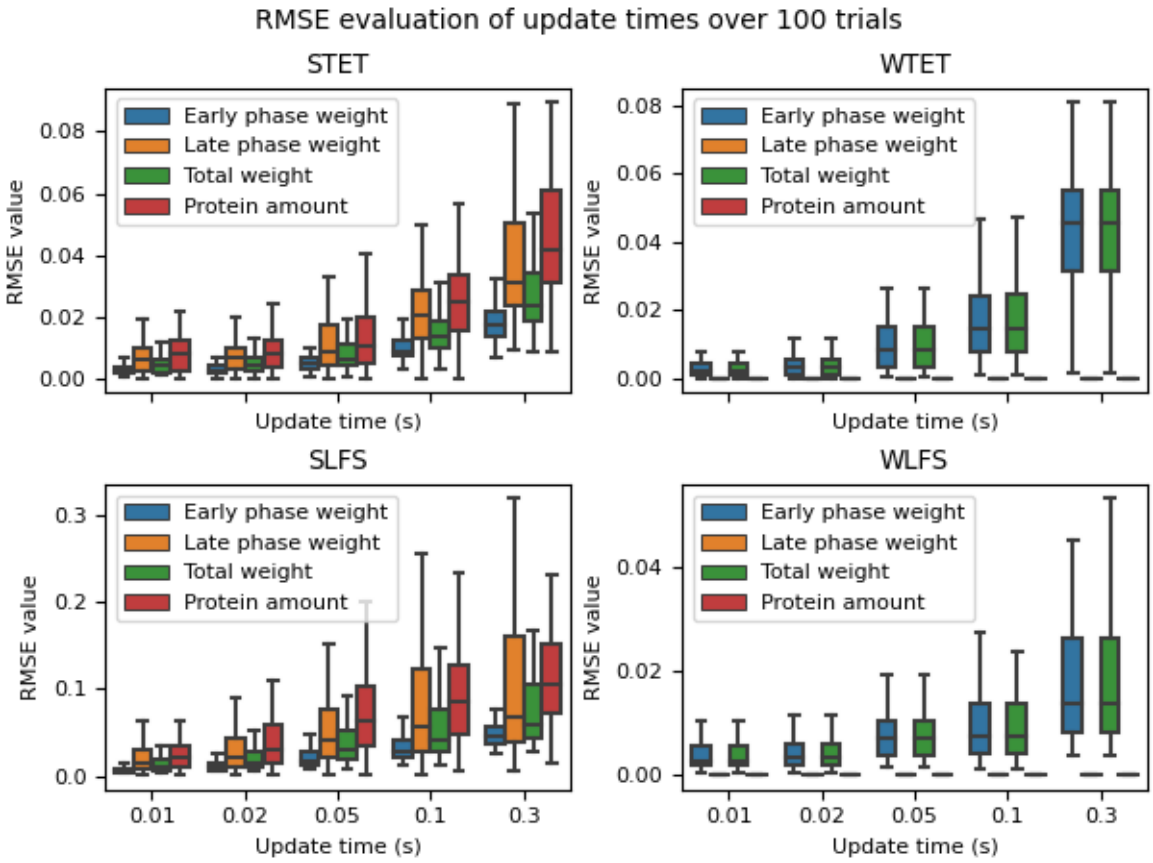

Figure 3.2: RMSE results obtained from updating protein amount and synaptic weights at slow timescales. The simulations are carried out for 100 trials. The unit of RMSE for synaptic weights is nC.

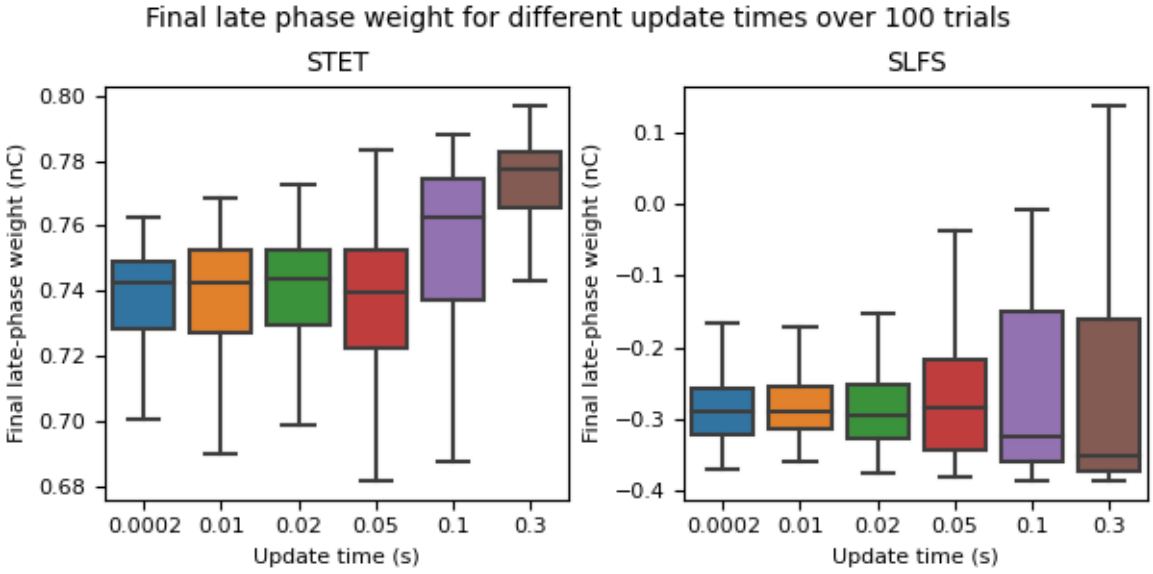

Figure 3.3: Final late-phase weight for the STET and SLFS protocols at different timescales over 100 trials. Note that the update time of 0.0002 s is the base timescale to be compared with slower timescales.

Looking at the time constants of our physical system in table 2.1, the base time step is sufficient for the evolution of all the system dynamics. This allows us to consider

that the variability in synaptic weights (fig. 3.1) is attributed to the variability in spike timings especially for the WTET and SLFS protocols. In these two protocols, the stimulation frequency is high relative to the stimulation duration, so the average number of spikes is low, but the spikes can also vary in number and timing resulting in a variability of weight values. For example, the SLFS protocol involves a stimulation frequency of 20 Hz over a stimulation duration of 0.15 s. On average, this results in 3 spikes per stimulation duration. However, the resulting spikes would range from 0 up to a number larger than 3 spikes for some trials. Even for a low number of spikes, the timing between spikes differs, which affects the calcium concentration and hence the synaptic weights. We can do a similar analysis for the WTET protocol which involves a stimulation of 100 Hz over a duration of 0.2 s. On average, this results in 20 spikes per stimulation duration, but the resulting spikes could vary between 0 to a number higher than 20 for some trials. For the STET protocol, the number of spikes is high, with an average of 100 spikes per stimulation duration. No matter the differences, the high number of spikes that occur close to eachother within a duration of 1 s produces a robust behavior in synaptic weights.

The effect of spike-time variability appears more significantly when updating the synaptic weights at slower timescales. The SLFS protocol has the highest RMSE values among all protocols (fig. 3.2), which can be attributed to the variability in spike times to a great extent. This is further verified by the final value of late-phase weight that maintains a high but consistent IQR and a consistent median up to 20 ms (fig. 3.3). The high error values can also be explained by the decay in calcium concentration during the update-time interval especially if the number of spikes is lower than the average. A similar reasoning can be applied for the early-phase weight error in the weak protocols which possess a low number of spikes, noting that with the slow update times, the behavior of the weak protocols is still correct in terms of no late-phase plasticity.

Combining the results from all protocols, update times up to 20 ms seem to be reasonable for the weight updates and can account for the time needed for the digital circuit to perform computations and PPU write instructions as well as CADC read instructions. However, we cannot judge how acceptable the resulting weights are unless we simulate the full network and have a clear idea about its behavior in terms of learning and memory recall with the modified plasticity rule. It is important to note, however, that the learning and recall protocols in the network use a stimulation frequency of 100 Hz over stimulation duration of 0.1 s and a break duration of 0.4 s. This stimulation scheme resembles that of the STET protocol, but the number of spikes and spike variability during a stimulation duration resembles that of the WTET protocol which is stimulated over 0.2 s. More simulation should be done in this sense to study the effect of spike variability at slow update times for the network stimulation and recall.

4.2 Outlook

For emulating the single-synapse plasticity protocols on BrainScaleS-2, further steps should be completed. The calcium concentration which links the spike occurrences

to the weights update will be computed by the analog circuit through exploiting special hardware features related to the AdEx model. Once the calcium concentration traces are computed on the analog core, they will be sampled for the weight updates. The simulations in this work were performed based on regular sampling of the calcium trace, but a stochastic sampling can be also investigated to improve performance. To emulate memory consolidation as faithfully as possible and obtain the expected biological behavior, the hardware parameters should be fine-tuned to the biological parameters. This requires mapping the total synaptic weight to an adequate range, tuning the adaptation module for the calcium concentration trace, and possibly tuning the neuron parameters such as its firing threshold. The corresponding performance can then be assessed by the resulting postsynaptic spikes and the obtained synaptic weights.

The behavior of the single-synapse emulation on BrainScaleS-2 would provide insights for emulating the full network presented in Luboeinski and Tetzlaff (2021). For the network behavior, different measures are used for assessing the memory and recall performance, namely the mutual information of the neural activity during learning and recall, and the pattern completion coefficient. As per the simulation in Luboeinski and Tetzlaff (2021), the network requires 2000 neurons, but BrainScaleS-2 has 512 neurons. The bulk performance relies on a cell assembly whose size is 150 neurons, and the rest are control neurons. An immediate solution is to keep the cell-assembly size and decrease the number of control neurons since they would have low activity. Another solution is to scale the full-network size, that is the cell-assembly and control neurons. This imposes a challenge on the memory behavior of cell assembly which could be resolved by strengthening the learning and recall stimulations.

Conclusion

In this work, we set the first building block towards emulating memory consolidation in a recurrent neural network on the neuromorphic BrainScaleS-2 system. We studied the behavior of a single synapse that incorporates the STC hypothesis by simulating standard plasticity protocols that induce different plasticity mechanisms. We then introduced the hardware constraint, namely the update time of the digital circuit, and ran simulations for different timescales to study the corresponding effect on the synaptic weights. The simulation results showed acceptable errors in the synaptic weights up to certain timescales, but reliable assessment of errors would be performed with full-network simulations using memory performance measures. Further steps are needed for faithful emulation of single-synapse plasticity including analog computation of calcium concentration and mapping the biological parameters to hardware parameters. Emulating the full network is the second stage after understanding the single-synapse behavior on BrainScaleS-2. The network emulation imposes even more challenges in terms of update times, memory performance, and network size. Nevertheless, using BrainScaleS-2 seems promising as this platform is designed to maximize precision in reproducing numerical simulation results (Indiveri and Liu, 2015).

Bibliography

Billaudelle, S. (2022). *From transistors to learning systems: Circuits and algorithms for brain-inspired computing*. Phd dissertation, University of Heidelberg.

Brette, R. and Gerstner, W. (2005). Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. *J Neurophysiol*, 94(5):3637–3642.

Frey, U. and Morris, R. G. (1997). Synaptic tagging and long-term potentiation. *Nature*, 385(6616):533–536.

Indiveri, G. and Liu, S. (2015). Memory and information processing in neuromorphic systems. *CoRR*, abs/1506.03264.

Kong, Q., Siauw, T., and Bayen, A. M. (2021). *Python programming and numerical methods: A guide for engineers and scientists*. Academic Press, an imprint of Elsevier.

Lamprecht, R. and LeDoux, J. (2004). Structural plasticity and memory. *Nature Reviews Neuroscience*, 5(1):45–54.

Luboeinski, J. and Tetzlaff, C. (2021). Memory consolidation and improvement by synaptic tagging and capture in recurrent neural networks. *Communications Biology*, 4(1):275.

Nadel, L. and Land, C. (2000). Memory traces revisited. *Nature Reviews Neuroscience*, 1(3):209–212.

Pehle, C., Billaudelle, S., Cramer, B., Kaiser, J., Schreiber, K., Stradmann, Y., Weis, J., Leibfried, A., Müller, E., and Schemmel, J. (2022). The brainscales-2 accelerated neuromorphic system with hybrid plasticity.

Schuman, C. D., Kulkarni, S. R., Parsa, M., Mitchell, J. P., Date, P., and Kay, B. (2022). Opportunities for neuromorphic computing algorithms and applications. *Nature Computational Science*, 2(1):10–19.