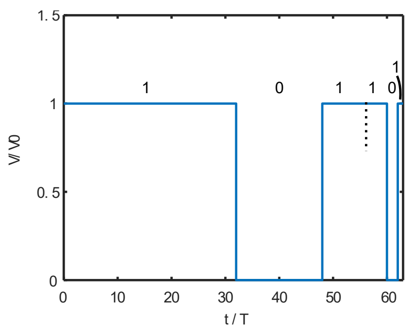
Solutions Exam EITP25 June 2022

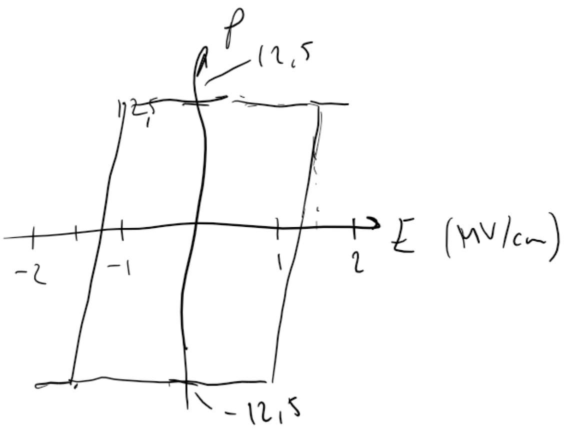
1. (3p) A) The elements of the input vector are encoded as an analog voltage amplitude applied to each row of a memristor crossbar array. The conductances of the memristors in the array encodes the elements of a matrix. The magnitude of the resulting current on each column of the array gives one element of the vector that results from a matrix-vector multiplication between the input vector and the matrix.
   Benefit: Matrix-vector multiplication in one time-step.
   Drawback: Requires the memristors to be linear (ohmic). + Requires DAC

   (3p) B) The elements of the input vector are encoded as a single voltage pulse of equal amplitude but a duration that scales with the magnitude of the vector element. The output current of each column is integrated, and the charge corresponds to an element of the matrix-vector multiplication.
   Benefit: Does not require Ohmic memristors + no DAC or ADC
   Drawback: Latency grows as 2^bits + precise pulse generator and integrator

   (4p) C) All bits are encoded as pulses having either the same amplitude V (in case of bit value 1) or 0 (in case of bit value 0). First pulse corresponds to the MSB and has the longest duration. If we define a shortest duration of 1, then the MSB is 2^5 long, the next is 2^4 long, and so on. The LSB has the duration 2^0=1.



2.a) (5p) The Polarisation current peak is nonzero between 1V --> 1.5 V, with thickness of 10 nm that means E between 1 MV/cm and 1.5 MV/cm. The height of the peak $I = 20\ \mu A$ and $dt = \frac{1}{4} * 10^{-4}\ s$. $Q = I * dt = 5 * 10^{-10}\ C$. $P = \frac{Q}{A} = 0.25 \frac{C}{m^2} = 25\ \mu C/cm^2$.



2 b) (5p) B is correct answer. In A (MIM with 1 nm oxide) the oxide is thin enough so that quantum tunneling is limiting, but there is not much modulation of the barrier width so the barrier will be almost equally transparent in both cases of polarization. In C, (MIS with 6 nm barrier) there is barrier modulation but the barrier thickness is too thick to give appreciable quantum tunneling, which means that the device is not sensitive to the barrier modulation. B (MIM with 4 nm) is a good middle ground, where the barrier is not too transparent for tunneling, and where barrier height modulation can have a good effect.

3 a) (2p) The GPU is a very parallelized architecture which is particularly good at parallelizing matrix-vector operations. Since the core operation in both inferencing and training in machine learning is the matrix-vector multiplication-accumulation operation it is well suited for this use-case.

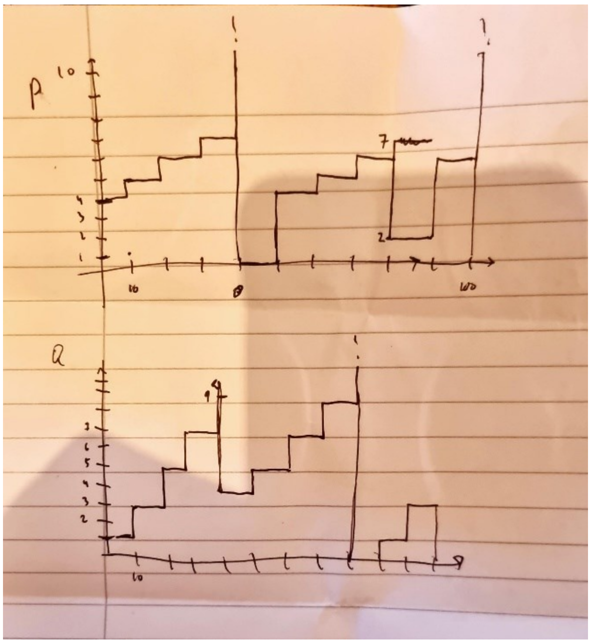
b) (3p) In-memory computing is solving the von-Neumann bottleneck problem which has to do with that in regular architectures memory needs to be read, transferred and rewritten to be operated on. It does this by utilizing fundamental laws of electronics (Ohm's law and Kirchhoff's law) to perform multiplication and addition directly where the memory is located, utilizing resistive memory units, preferably analog.

c) (3p) Case a is the most harmful as it is non-linear and asymmetric. In particular the asymmetry is problematic. Case b also adds device-to-device variation, but this is actually helping performance, counteracting the asymmetry problem. Case C has only non-linear behaviour which is not so problematic, and this case should thus be least harmful.

d) (2p) Negative weights can be represented in various ways. Pick one of these:

- Two parallel memristor devices that connect to the inputs of a comparator. The weight is then the difference between the conductances
- Use a column of reference devices programmed at mid-level conductance and subtract the current through this column from the result of any other columns in the array. In this way the lower half of conductances represents negative weights, upper half positive.

4. a) 6p



P: spikes at 40 and 100 ms, Q: spikes at 80 ms.

b) (4p) **The relative timing** between pre-neuron spikes and post-neuron spikes matters. Define $\Delta t = t_{post} - t_{pre}$ as the difference in time between these events on a synaptic connection. For $\Delta t > 0$ the synapse weight should be potentiated by an amount that is inversely proportional to $|\Delta t|$, usually $\Delta w_+ = Aexp(-|\Delta t|)$. Similarly, if $\Delta t < 0$ then the weight should be depressed in a similar way $\Delta w_- = -Bexp(-|\Delta t|)$. **Thus spikes close in time leads to larger weight change** than less correlated spikes. **The logic behind** this behavior is that preneuron spikes arriving before the postspikes carry relevant information and the connection should thus be strengthened, while spikes arriving after the postneuron spikes carry no relevant information and should be depressed.

5. a) (3p) Since these device technologies require a certain current density to switch **this has to be delivered through the access device**. Typically, the access device is a **MOSFET, which can drive a limited current per gate width**. And thus, it is often **the width of the MOSFET that limits amount of used area** for these devices.

b) (3p) Resistance contrast or Tunneling Magnetoresistance (TMR) in an MTJ is due to a **difference in tunneling probability** across an insulating barrier depending on whether the magnetization of the reference and free layer are parallel or antiparallel. **Electrons have to tunnel between states with the same spin quantum number.** If they are **parallel then there is a match between occupied states for a certain spin states on one side and the corresponding available states on the other side of the barrier**. If the magnetizations are antiparallel then there are fewer available states to tunnel through (or fewer to tunnel from).

c) (4p) The limited retentation in FeFET originates in the fact that the **semiconductor channel cannot efficiently screen the polarization charg**e, in particularly when depleted. Thus, a **depolarization field forms over the ferroelectric film** which leads to a **lower equilibrium polarization** charge, and thus loss of retention. A way to avoid this problem is to **connect a MFM capacitor in series with the MOS structure**, essentially having an ordinary MOSFET with a ferroelectric capacitor on top. This allows for better screening of the polarization charge and individual tuning of the MIS and MFM capacitances.

6.

a) (5p) $I_{sel} = V_{on}G_{off} = 1\,V * \frac{10\ \mu S}{100} = 0.1\ \mu A., I_{sneak} = N_{half}I_{on}(0.5\,V) = 2 * 99 * \frac{1\,V * G_{on}(1\,V)}{\eta} = 198 * \frac{10\ \mu A}{1000} \approx 2.0\ \mu A.$

b) (5 p)
PCM: Diodes – very non-linear, **high current levels**, **unidirectional** (works with PCM).
RRAM: Oxide threshold switches - **Bidirectional devices** (required for switching RRAM), highly nonlinear. **Too low current levels can be an issue**.