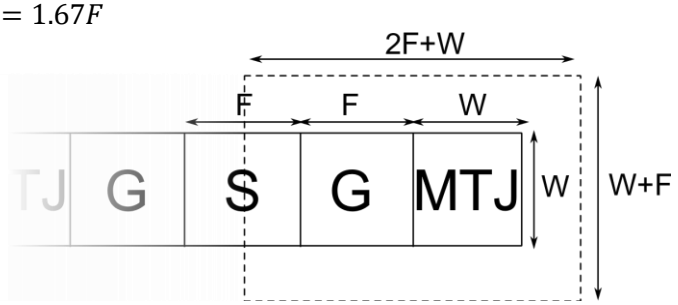
Solutions to the Written Exam EITP25 2021 3 June 2021

1.
   A) SRAM may be fast but it takes up too much space: >100F^2. (3p)
   B) NAND is slower because a whole string of devices is addressed at once. Even if all devices except for the target device will be in the on-state, their on-resistance will be higher than that of a metal line. In NOR the target transistor is addressed through metal lines, giving lower series resistance.
   C) The 3x3 kernel with stride of one means that each input data point transferred will be used in 3x3=9 floating point calculations. Thus the flops/byte = 9: 1000 Gbyte/s → With 9 flop/byte → 9 Teraflop performance only 9% of the specified performance. The limiting factor here is the memory transfer. By increasing the kernel size one would further increase the flops/byte.
2.
   A) Represent the transpose of the array as the conductance of the PCM devices in the memory array. Apply the vector elements x_i as $V_i = x_i V_0$ on each row, where V_0 is some known voltage. Read out the current on each column, it will be proportional to the dot product of one row of the array and the vector. For the first column one would get: $I_1 = 3V_0 * 4 + 2V_0 * 2 + 1V_0 * 1 = (12 + 4 + 1)V_0 = 17V_0$. The result of the full multiplication is $[17 \quad 17 \quad 30]^T$.
   B) Neurons 1 and 2 both have spikes very close in time before the spike from neuron 4. Thus, the synapses between 1-4 and 2-4 should be potentiated, 2-4 more than 1-4 because it has two spikes in close succession. Neuron 3 has a spike shortly after the spike from 4, and the synapse 3-4 will therefore be depressed. The spikes further away from spike 4 will have minimal impact on the change of weights.
   C) A LIF-neuron implements an integrating membrane potential on which charge can be stored. The potential also leaks down to an equilibrium value over time. If the potential reaches an upper threshold an output spike is generated and the potential is reset. The integrating membrane can be realized by a capacitor, the leakage by a resistor in parallel, and the threshold by a comparator.
3.
   A) Tunneling magnetoresistance (TER) is caused by different transmission probability through a magnetic tunnel junction, a junction of two magnetic materials separated by an insulator, depending on whether the two layers are magnetized in the same direction (parallel) or opposite (anti-parallel), because electrons must tunnel between states with the same spin orientation. Giant TER is obtained by using a crystalline insulator MgO in which transmission of non-spinpolarised electronic states is much reduced.
   B) Generally: $I_{max} = \frac{2mA}{\mu m} * F \rightarrow J_{max} = \frac{I_{max}}{F*F} = \frac{20\frac{A}{cm}*\cancel{F}}{\cancel{F}*F} = J_c = 3\frac{MA}{cm^2} \rightarrow F = \frac{20}{3E6} = 67\ nm = 1.67F$



➔ $A_{MRAM} = (2 + 1.67)(1 + 1.67)F^2 = 9.8F^2$.

In this case: First test whether Imax is large enough with W = 1F! $I_{max} = \frac{2mA}{\mu m} * 0.04\mu m = 80\ \mu A > J_c * F^2 = \frac{3MA}{cm^2} * (40 * 10^{-7})^2 = 48\ \mu A$. So no scaling is actually needed, the MOSFET is sufficient for $6F^2$ device size.

(As this is a bit tricky and easy to miss, I decided to give full credits to both solutions)

C) For low write currents, the magnetization switching in STT-MRAM is stochastic. One can utilize this stochasticity to realize STDP learning in which a pre-pulse coming just before the post-pulse can give probability for potentiation of the synapse, and a pulse coming after gives a probability for depression, mimicking the shape of the regular STDP curve. Upon potentiation and depression the weight is changed to either the maximum or minimum value.

4.

A) P_r = 15 µC/cm², E_c^+ = 1 MV/cm, E_c^- = -0.5 MV/cm → $E_c = \frac{E_c^+ - E_c^-}{2} = 0.75 \frac{MV}{cm}$.

B) $E_c = 0.75 \frac{MV}{cm}, t_{HZO} = 5\ nm \rightarrow V_c = 0.375\ V$

$$r = 25\ nm \rightarrow A = \pi r^2 = 2 * 10^{-15} m^2,$$
$$\varepsilon_{HZO} = 25\varepsilon_0$$
$$C_{cap} = \frac{\varepsilon_{HZO} A}{t_{HZO}} = 0.09\ fF$$
$$\rightarrow Q_{cap} = C_{cap} * V_c = 3.3 * 10^{-17}\ C$$
$$P_r = 15\ \frac{\mu C}{cm^2} = 0.15 \frac{C}{m^2} \rightarrow Q_{FE} = 0.15A = 3 * 10^{-16} C$$
$$Energy = \left(Q_{cap} + 2 * Q_{FE}\right) * \frac{V_c}{2} = 0.12\ fJ$$

Note: $2Q_{FE}$ because you change the polarisation state by $2*P_r$ ($-P_r$ → $+P_r$ for example).

C) FeFETs are essentially a MOSFET in which the gate dielectric contains a ferroelectric film. The polarisation state of the ferroelectric film will shift the threshold voltage of the MOSFET either negatively or positively. This is written by pulsing either a large negative or positive gate voltage. At an applied gate voltage in between the two threshold voltages thus allows to read a source/drain current which corresponds either to the off- or on-state of the transistor depending on the polarisation state. This can give a very large resistance contrast between the two states. In the off state, the depleted semiconductor causes the polarisation charge to be uncompensated on one side of the capacitor which leads to a depolarisation field which can reduce the effective polarisation charge and thus preventing long retention times. Endurance can be an issue since upon voltage cycling the generation of defects at the semiconductor-ferroelectric interface will eventually screen the polarisation charge and thus leading to a subsequent shrinking of the polarisation-induced threshold voltage shift.

5.

A) The purpose of the selector is to make the current-voltage characteristic of the memory cell non-linear so that the leakage through half-selected cells on the same row or column is sufficiently low.

B) P-n junctions are good because they are two-terminal devices that can be vertical devices that can be made directly below or on top of the memory device, thus not taking up additional chip area. Their nonlinearity can be high and current density is high. They only work for memory devices that are unipolar such as PCMs because they are rectifying.

C) In the worst case, the selected device is in the off-state and all other devices are in the on state. With V/2 biasing scheme all unselected devices have no bias across them thus providing no current, while half-selected devices have V/2 bias. Thus $I_{half} = 10^{-15}\exp(25 * 0.5) = 0.27\ nA. I_{selected} = 10^{-16}\exp(26 * 1) = 20\ \mu A$. In a 1000x1000 array, there are 999 + 999 halfselected devices which give $SNR = \frac{20\ \mu A}{1998 * 0.27\ nA} = 37$