Solutions to Reexam EITP25 18 Aug 2021

1. A: DRAM consists of one MOSFET with drain contact connected to a capacitor which charge state corresponds to the memory state. The source contact is connected to the bit line and gate contact to the word line. Writing to DRAM means to pre-charge the bit-line to the desired state and then selecting the appropriate wordline to open the MOSFET channel, which then writes the bitline potential to the capacitor. DRAM is not used for solid state drives because it is non-volatile, i.e. it's capacitor charge leaks out and needs constant refreshing.

   B. Eb > 1.4 eV thus emission above the barrier is not limiting retention. Thus tunnelling could be an issue: $t_T > 10\ years * 365\ days * 24\ h * 60\ min * 60\ s = 3 * 10^8 s$

   $$t_T = \frac{1}{L^2 f_0^*} \exp\left(\frac{2\sqrt{2m^* E_b}}{\hbar} t_{ox}\right) \rightarrow t_{ox} = \frac{\ln(L^2 f_0^* t_T)\,\hbar}{2\sqrt{2m^* E_b}} = 3.8\ nm$$

   C. The most important mathematical operation in machine learning is the matrix vector multiplication accumulation (MAC) operation. The GPU has thousands of computing cores that can perform matrix-vector multiplication in parallel. This makes it suitable for ML.

   D. For a small CNN, it is the memory transfer rate that limits the performance, because for each data block the computation of the network is quick and can be done in few steps, the hardware will need to wait for the data to arrive through the limited I/O port.

2. A: These are the steps:

   - Forward propagation step: Propagate data from the input layer onto the output layer generating the weighted input and activation of each layer
   - Compute the output error: Calculate the error/cost and gradient of the cost with respect to the weights for the output layer
   - Backpropagate the error from output to input layer to calculate the gradient of the cost with respect to the weights for each layer using the information from the forward pass and the previous backpropagation steps.
   - Update the weights in the direction of negative cost gradient with a step length given by a learning rate.

   B. A convolutional layer in a CNN consists of a number of filters with a size smaller than the input data that are mapped across the input data generating a feature map where each pixel is defined by the weighted input from a particular portion of the data. The feature map is often condensed using a pooling layer. A CNN is built up by a combination of convolutional, pooling layers as well as fully connected layers into a full network. Benefit 1: A CNN makes use of spatial relationships in the data, (ex: neighboring pixels form a line). Benefit 2: A CNN uses less weights per filter compared to a fully connected layer, which helps with parallellization on GPU.

   C. Lateral inhibition is when a neuron in a layer of a spiking neural network spikes and then prevents other neurons in the same layer from firing for some time. This can be either by reducing their membrane potential somewhat (soft inhibition) or by completely resetting their potential (winner-takes-all). Lateral inhibition promotes learning by forcing neurons to specialize on not already learned patterns.

3. A. STT is the effect that a spin-polarized electron current will induce a torque on the magnetization of a magnetic layer, so that there is a force for the spins to align. If the current is strong enough the magnetization in the free layer can be flipped.

   B. MRAM requires less energy to write than PCM. PCM write requires a long enough time for crystallization to occur, thus limiting how short a write pulse can be. Also, the current needs to be high enough to reach above the crystallization temperature. (STT-)MRAM has no such limitation. A strong enough current pulse can lead to a state change regardless of the timing, thus leading to lower energy consumption.

   C. A 2-PCM synapse consists of two PCM devices connected to a comparator. One PCM represents a positive weight part and the other a negative part. The real weight is the difference in conductance between the two devices. This allows for representation of negative weights as well as to mostly use the less costly SET operation and avoiding the abrupt RESET operation, which is only performed once one of the two devices reaches the maximum conductance state.

4. A: PUND consists of four triangular voltage pulses, two positive followed by two negative. The first P-pulse will program the ferroelectric polarization into the positive direction, resulting in a current that depends on the amount of ferroelectric charge programmed in addition to other currents. The current corresponding to the second positive U-pulse will NOT include any ferroelectric current, as all polarization is already set in this direction. Thus one can subtract the current of the U-pulse from the P-pulse, resulting in a near isolation of the ferroelectric contribution. Similarly, the N-pulse will program the polarization to the other direction followed by the D-pulse for the other contributions. N-D gives the polarization current contribution in the negative direction. Finally one integrates the currents to obtain polarization charge instead.

   B. FeFETs can have issues with retention due to the depolarization field that occurs due to incomplete screening of the polarization charge in the depleted semiconductor channel. If the depolarization field is larger than the coercive field it leads to partial depolarization which impacts retention. This can be avoided by using a MFMIS structure in which the ferroelectric charge is screened by metal on both sides.

   C. The way to achieve very high Roff/Ron ratios with FTJs is to have a semiconductor on one contact. In this way the semiconductor will be partly depleted in one polarization state leading to a significantly longer tunnelling distance which greatly increases R_{off} if the barrier is thin enough so that tunnelling limits the transport. In a regular MIM FTJ one only modulates the height of the barrier leading to highest Roff/Ron when thermionic emission dominates the current transport.

5. A. In electrochemical metallization metal ions from one contact is ionized and diffuses to the other contact where it is neutralized and builds a metallic filament. The on-state is reached once the filament reaches across the oxide gap (or close enough for tunnelling). In valence change memory on the other hand, oxygen vacancies are formed in the oxide (near the active contact). These build up a conductive filament.

   B. Ron is determined by the width/conductance of the filament. By allowing a wider/stronger filament to form by choosing a higher current compliance level upon SET one can control Ron.

C. Roff is determined by the size of the gap upon breaking the filament. Variation can happen because the active metal is oxidized in an uncontrolled fashion. By creating a controlled oxide at the active metal surface a reproducible Roff can be achieved.

D. A bipolar ReRAM needs to be able to have high current in both voltage polarities, since SET and RESET are done with opposite polarities. Thus, diode selectors are out of question. Although oxide Schottky barrier selectors are bipolar the current level may not be sufficient for SETting ReRAM. MOSFETs appear to be the best choice for a selector as they can provide high current with both biases and can be highly non-linear thanks to a high Ion/Ioff ratio.