Reading #3 Quiz

Started: Feb 10 at 6:07pm

Quiz Instructions

iii

This quiz is based on domain_specific_hw_accelerators.pdf (https://canvas.asu.edu/courses/212341/files/102624391?wrap=1)

Question 1 1 pts

ASICs provide higher efficiency than GPUs for domain-specific acceleration but come at the cost of higher nonrecurring engineering (NRE) costs.



True

 \bigcirc

False

::

Question 2 1 pts

In modern technology nodes, the energy required for moving data (such as reading from memory) and instruction processing (such as fetching an instruction) is significantly lower than performing a simple arithmetic operation.

 \bigcirc

True



False

Question 3 1 pts

Which of the following is NOT a major technique used by domain-specific accelerators to achieve performance gains?

 \bigcirc

Data specialization



Parallelism



Increased clock speeds

0

Optimized memory

The likelihood that

The likelihood that genomic alignment algorithms will evolve over time

The need for higher single-threaded performance

Question 7 1 pts

Which of the following design strategies would BEST ensure that a domain-specific accelerator remains useful even as new algorithms emerge?

O Implementing fixed-function hardware with no software programmability

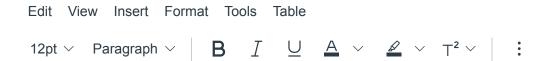
Using a GPU to maintain generality



Adding specialized instructions to a programmable processor

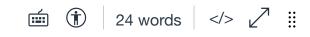
Question 8 1 pts

Your company is designing a DSA for natural language processing (NLP) tasks. Given the large memory footprint of transformer-based models, how would you optimize your accelerator to handle such workloads efficiently? Write a few techniques you'd apply. Answer in no more than 25 words.



I would apply model pruning, quantization, mixed-precision arithmetic, memory pooling, and layer fusion to reduce memory usage and enhance performance in transformer-based NLP tasks.

p **∷**



Question 9 1 pts

You are designing a DSA for real-time speech recognition (such as one in Google Home or Alexa). This accelerator must process continuous audio streams with low latency. What architectural decisions would you take? Write a few techniques you'd apply. Answer in no more than 25 words.

2/10/25, 6:18 PM Quiz: Reading #3 Quiz

Edit View Insert Format Tools Table

12pt \vee Paragraph \vee B I $\underline{\cup}$ \underline{A} \vee $\underline{\mathscr{D}}$ \vee \top^2 \vee :

I'd use pipelining, stream-based processing, low-latency neural networks, real-time data buffering, and hardware-level acceleration with custom DSPs to ensure fast, continuous audio processing.

p :: Question 10 1 pts

List names of the some DSAs mentioned in the paper.

Edit View Insert Format Tools Table

12pt \vee Paragraph \vee B I \cup A \vee \nearrow \vee $\top^2 \vee$:

The List of DSAs:

Darwin

Efficient Inference Engine

Sparse Convolutional Neural Network Accelerator

NVIDIA Deep Learning Accelerator

Google Tensor Processing Unit

CHARM

MARS

MSE

2/10/25, 6:18 PM Quiz: Reading #3 Quiz

р

(†) 24 words </> √/ ∷

No new data to save. Last checked at 6:18pm

Submit Quiz