QUESTÕES – Domain Specific Hardware Accelerators (DSA)

1. O artigo inicia com a frase "...however, the energy required to fetch and interpret an instruction is 10× to 4000× more than that required to perform a simple operation such as ADD". Isto no leva a concluir que CPUs são extremamente ineficientes.
    a. Por que ocorre esta ineficiência?
    b. Por nas últimas décadas esta ineficiência foi (continua) a ser aceita?

2. Como os autores define DSA, e por que justificam sua adoção?

3. Interprete a seguinte frase "...the application must be **reworked**, codesigning the application with the accelerator, to reduce memory bandwidth and memory footprint"

4. Fontes de aceleração.

    a. **Data specialization.** Qual o exemplo apontado? Na frase "In many cases, specialized logic can perform the entire inner loop in a single cycle", o que o Autor diz necessário à arquitetura?

    b. Comente os dados abaixo, relacionados ao tópico "Data specialization."

	Cycles	Energy
an Intel Xeon E5-2620 4-issue, out-of-order 14nm CPU	38	81 nJ
40 nm Darwin accelerator	1	3.1pJ (0.3 pJ comp/ 2.8pJ mem)

	Energy
32-b integer add	63 fJ in 28nm CMOS
40 nm Darwin accelerator	250 pJ 28nm ARM A-15

    c. Pesquise o que é "compressed- sparse-column (CSC) format". Como o Autor aponta que utilizar este formato auxilia aulixa na redução de energia?

    d. **Paralelismo**. Qual o mecanismo chave apontado para obter-se paralelismo efetivo?

    e. Cite 3 características que o Autor menciona serem necessárias para garantir efetivo paralelismo com dezenas de elementos de processamento?

    f. Na pág. 51 o Autor menciona "With double buffering of the inputs and outputs, the arrays are working continuously....". Pesquise o que seria esta técnica "double buffering" e apresente exemplos.

    g. **Local and optimized memory**. Qual a arquitetura de memória apontada? Cite exemplo de técnica apontada para melhorar a lagura de banda e quais sistemas o utilizam.

    h. Porque o autor afirma que o uso de um sistema de memória, mesmo com vários níveis de memória cache limita o paralelismo?

i. Reduced overhead. O que seria este item?

j. Explique a instrução matrix- multiply-accumulate instruction (HMMA) da NVIDIA.

5. Página 52 - Acceleration Options – comente a figura 2, apontando os compromissos citados no texto.

6. Qual o significado de "codesign" no artigo? Porque ele é necessário?

7. Comente a tabela 2, relativo ao breakdown de power entre lógica e memória.

8. (IMPORTANTE) Comente a frase "One approach to building accelerators for broad domains is to add specialized instructions to a general-purpose processor." Segundo o autor compensa um acelerador para uma aplicação específica? Por quê? (praticamente toda a página 54).

Total Cost of Ownership (TCO) – custo do acelerador. Questões de nodo tecnológico, volume e custo de engenharia (NRE) – figura 1

9. Comente os dados apresentados no texto relativo à área e energia, resumidos na tabela abaixo.

	Area	Energy
14 nm technology, arithmetic – 8bit	4 μm^2	10 fJ
14 nm technology, arithmetic – PF, precisão dupla	3600 μm^2	5 pJ
Memória SRAM 8KB – 14nm	0.013 μm^2/bit	50 fJ/bit
Comunicação com a Memória SRAM		100fJ/bit-mm
LPDDR4 memory		4 pJ/bit
High-speed off-chip channels - SerDes		10pJ/bit

Communication energy remains roughly constant. This nonuniform scaling makes communication—such as nonlocal memory access—even more critical in future systems.

10. Apresente um exemplo de "Placing memories on interposers can give bandwidths up to 1TB/s, but at the expense of limited capacity."

Ao final:*Ultima tely, we expect that computer science curricula will evolve to teach algorithms and complexity with a cost model that more accurately reflects the reality of modern computing hardware.*