

QUESTÕES – Domain Specific Hardware Accelerators (DSA)

Autor: Fernando Gehm Moraes

Revisão: Ney Laert Vilar Calazans

Última versão de 17/05/2022, 17:53

- 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.
 - Por que ocorre esta ineficiência?
 - Por que nas últimas décadas esta ineficiência foi (continua) a ser aceita?
- Como os autores definem DSA, e como justificam sua adoção?
- Interprete a seguinte frase "...the application must be reworked, codesigning the application with the accelerator, to reduce memory bandwidth and memory footprint"
- Fontes de aceleração.
 - 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?
 - 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 / 2.8 pJ mem)

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

- Pesquise o que é "compressed- sparse-column (CSC) format". Como o Autor aponta que utilizar este formato auxilia na redução de energia?
- Paralelismo.** Qual o mecanismo chave apontado para se obter paralelismo efetivo?
- Cite 3 características que o Autor menciona serem necessárias para garantir paralelismo efetivo com dezenas de elementos de processamento.
- Na pág. 51 os Autores mencionam "With double buffering of the inputs and outputs, the arrays are working continuously....". Pesquise o que seria esta técnica "double buffering" e apresente exemplos.
- Local and optimized memory.** Qual a arquitetura de memória apontada? Cite exemplo de técnica apontada para melhorar a largura de banda e que sistemas a utilizam.
- Porque os Autores afirmam que o uso de um sistema de memória, mesmo com vários níveis de memória cache limita o paralelismo?
- Reduced overhead. O que significa?
- Explique a instrução matrix- multiply-accumulate instruction (HMMA) da NVIDIA.

5. Página 52 - Acceleration Options – comente a Figura 2 do artigo, apontando os compromissos citados no texto.
6. Qual o significado de “codesign” no artigo? Por que ele é necessário?
7. Comente a Tabela 2 do artigo, relativa 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
14nm technology, arithmetic – 8bit	4 μm^2	10 fJ
14nm technology, arithmetic – PF, precisão dupla	3600 μm^2	5 pJ
Memória SRAM 8KB – 14nm	0.013 $\mu\text{m}^2/\text{bit}$	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. 10. Apresente um exemplo de trabalho que justifique a frase “Placing memories on interposers can give bandwidths up to 1TB/s, but at the expense of limited capacity.”

Ao final: *Ultimately, 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.*