PROJETO DE HARDWARE LOW-POWER PARA IDCT DO PADRÃO HEVC

JONES GÖEBEL; RUHAN CONCEIÇÃO; LUCIANO AGOSTINI; BRUNO ZATT; MARCELO PORTO

Universidade Federal de Pelotas - Grupo de Arquitetura e Circuitos Integrados
{jwgoebel, radconceicao, agostini, zatt, porto}@inf.ufpel.edu.br

1. INTRODUÇÃO

Atualmente, os vídeos digitais são cada vez mais utilizados em dispositivos, como smartphones, televisores, Blu-ray players, etc. Contudo, nesses dispositivos é necessária a utilização de *codecs* (codificadores/decodificadores) de vídeo, visando reduzir a quantidade de dados necessários para a representação dos vídeos digitais. Destaca-se que o processo de codificação/decodificação deve ser realizado de forma que os dispositivos que implementam o codec sejam capazes de armazenar e transmitir os vídeos em tempo real.

O aumento das resoluções de vídeo se tornou um dos principais desafios da área de pesquisa em codificação de vídeo, pois implica em um processamento de mais dados. Este fato torna a codificação/decodificação uma tarefa infactível para os processadores de proposito gerais. Para solucionar este problema, umas das alternativas é a implementação de arquiteturas de hardware dedicadas para os módulos do codec de vídeo.

O padrão estado da arte em codificação de vídeo é o *High Efficiency Video Coding* - HEVC (JCT-VC, 2013), que atinge uma taxa de compressão superior a 50% em relação ao seu antecessor, o H.264/AVC (ITU, 2005), mantendo a mesma qualidade do vídeo. Para que o HEVC atingisse essa taxa de compressão, teve que haver um aumento na complexidade computacional (BOSSEN, 2012).

O princípio básico da compressão de vídeos digitais consiste na exploração das informações redundantes. As redundâncias encontradas nos vídeos podem ser classificadas como: espacial, temporal e entrópica. A redundância espacial ocorre pela similaridade de uma amostra com seus vizinhos dentro um mesmo quadro do vídeo, sendo ela explorada pela predição intra-quadro dentro do HEVC. A redundância temporal é explorada pela predição inter-quadros que é formada a partir de informações replicadas entre os quadros vizinhos que compõe o vídeo (AGOSTINI,2007). Assim, é possível utilizar apenas a diferença entre os dados preditos com os originais, os quais são chamados de resíduos. Estes resíduos são manipulados pelas etapas chamadas de transformada, quantização e entropia.

Após o processamento de todas estas etapas acima mencionadas, o quadro deve ser reconstruído novamente visando ser utilizado nas próximas predições como referência. Desta forma, o codificador também implementa um decodificador, sendo composto (entre outros) pelos módulos da quantização inversa e transformada inversa. A transformada inversa apresenta vários desafios, pelo fato de ser implementada tanto no codificador quanto no decodificador, possuindo requisitos de desempenho bem distinto nestes dois cenários.

Este trabalho apresenta um projeto de hardware proposto para a Transformada Discreta dos Cossenos Inversa (*Inverse Discrete Cosine* Transform – IDCT), principal transformada do módulo da transformada inversa do HEVC. A arquitetura é capaz de processar todos os tamanhos de IDCT estipulados no HEVC (4x4, 8x8, 16x16 e 32x32) com processamento fixo de 32 amostras por ciclo. Além disso, a arquitetura tem a capacidade de manusear TUs heterogenias

(TU – estrutura utilizada para representar um bloco de pixels da imagem na etapa da transformada), permitindo o processamento de diferentes tamanhos de transformada ao mesmo tempo. A arquitetura foi sintetizada visando atingir o desempenho necessário para processar vídeos com a resolução 8K (7680x4320 pixels) à taxa de 60 quadros por segundo. Além disso, o projeto de hardware foi realizado visando propiciar à arquitetura desenvolvida uma baixa dissipação de potência, visto que o processo de codificação é amplamente executado em dispositivos embarcados, os quais apresentam restrições energéticas mais severas.

2. METODOLOGIA

A IDCT – assim como a DCT – utilizada no HEVC apresenta algumas peculiaridades tal como a propriedade da separabilidade (GHANBARI, 2003). Este princípio consiste em aplicar duas IDCT-1D sucessivamente, interligadas por uma etapa de transposição, visando obter a IDCT-2D. A segunda característica destacada é o fato das transformadas de tamanho menor compõem transformadas de tamanho maior. Desta forma, a transformada de 4 pontos está inserida dentro da transformada de 8 pontos, que por sua vez esta inserida na transformada de 16 pontos e assim por diante.

A arquitetura IDCT-2D apresentada neste trabalho é capaz de processar diversos tamanhos de transformada, mantendo o processamento fixo de 32 amostras por ciclo. Tendo em vista que a arquitetura foi desenvolvida visando apresentar uma baixa dissipação de potência, empregando registradores *clock-gating* (PAL, 2014) no projeto, os quais reduzem a potência dinâmica em regiões ociosas do hardware apenas controlando a arvore de clock.

A Figura 1 ilustra um diagrama de blocos da arquitetura IDCT-1D multi-tamanho proposta. Esta contém oito módulos principais de processamento de dados (PD) e dois módulos de distribuição de dados para os módulos PDs. O Primeiro PD é capaz de processar IDCTs de tamanho de 4, 8, 16 e 32 pontos por ciclo, sendo chamada de PD-32p. Seguindo a mesma ideia foi desenvolvido uma PD-16p, duas PD-8p e finalmente quatro PD-4p.

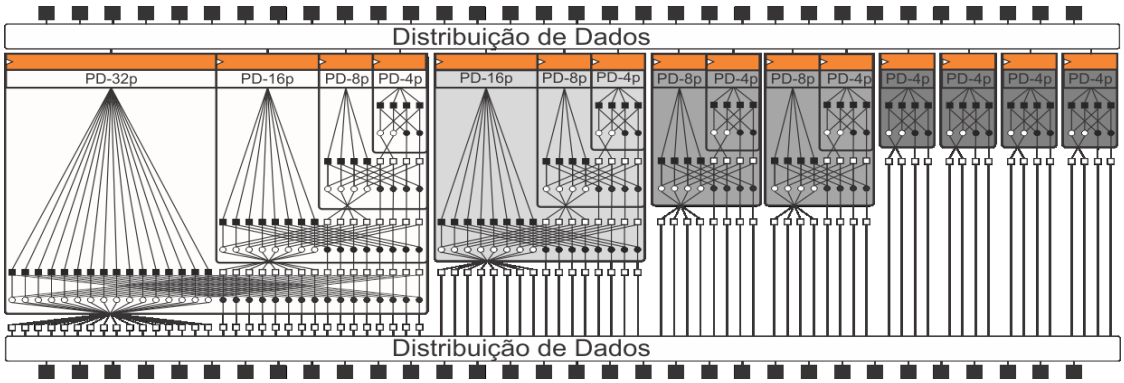

Figura 1. Diagrama de blocos da arquitetura desenvolvida.

Quando a arquitetura está processando IDCT 32 pontos, o modulo PD-32p ira consumir 32 amostras por ciclo, enquanto os demais PDs estarão ociosos. Entretanto, quando a arquitetura processa duas IDCT de 16 pontos, o modulo PD-32p consome 16 amostras por ciclo e o modulo PD-16p consome as outras 16 amostras restantes. Quando a arquitetura estive processando apenas IDCTs de 4-pontos, cada modulo PD-4p ira consumir 4 amostras por ciclo e os outros módulos consumirão, cada um, 4 amostras por ciclo. Esta característica garante que a arquitetura mantenha consumo fixo de 32 amostras por ciclo independente da configuração de blocos a ser processado.

Podemos notar que quando a arquitetura está processando apenas IDCT de 32-pontos, ela utiliza apenas o modulo PD-32p enquanto que os demais módulos não são utilizados no mesmo instante de tempo. Visando reduzir o chaveamento nos módulos ociosos, implementou-se registradores *clock-gating* na entrada de todos os PDs, evitando assim o chaveamento provocado pelas entrada dos dados. Na Figura 1, os registradores *clock-gating* são ilustrados através de caixas laranjas (CG). Os flip-flops utilizados para realizar o *clock-gating* são controlados pela unidade de controle que não aparece na Figura.

Finalmente, o projeto de hardware desenvolvido também é composto por um buffer de transposição, o qual realiza a interconexão entre a primeira e a segunda IDCT-1D. Em outras palavras, o buffer é responsável em receber os dados processados pela primeira IDCT-1D e armazená-los em coluna, repassando-os para a segunda IDCT-1D em linha. A matriz de transposição possui a mesma estrutura da matriz em outros trabalhos, mas a diferença entre esta e as demais encontra-se no controle, o qual permite que a arquitetura posa manusear todos os tipos de TUs na transposição da matriz.

3. RESULTADOS E DISCUSSÃO

A arquitetura desenvolvida foi descrita em VHDL e sintetizada em ASIC utilizando ferramenta da Cadence Encounter RTL Compiler sendo simulada utilizando a biblioteca *standard-cells* de 45nm da Nangate para 0,95V. A arquitetura foi validada utilizando entradas reais da IDCT utilizada no software de referência do HEVC (HM 16.7).

Para poder observar o ganho em termos de redução de dissipação de potência proporcionado pela técnica de *clock-gating*, foi realizada a descrição de duas arquiteturas: a primeira implementando tal técnica e a segunda não. Os resultados obtidos na síntese podem ser observados na Tabela 1. Para a geração dos resultados foi utilizado a frequência de 93,3 MHz.

Tabela.1 – Resultados de Síntese.

Arquitetura	Área (K gates)	Amostras por ciclo	Frequência (MHz)	Power total (mW)
Sem CG	381,7	32	93,3	88,85
Com CG	382,3	32	93,3	30,02

Como pode ser observado na Tabela 1, a arquitetura com *clock-gating* obteve uma redução de 66,2% na dissipação de potência. Resultados de síntese também demonstraram que a arquitetura com *clock-gating* usou 382,3k portas lógicas (305.054,65 μm²) e a sem a *clock-gating* usou 381,7K portas lógicas (304.600,59 μm²) o que representa um aumento de apenas 0,2% em área. Destaca-se que a contagem das portas lógicas é realizada baseada na NAND de duas entradas (0,798 μm²).

Utilizando a frequência de 93,3MHz a arquitetura pode atingir o processamento a 60 quadros por segundo de vídeos com resoluções UHD 8K e 240 quadros por segundo em vídeos UHD 4K (3840x2160 pixels).

Na Tabela 2 são apresentados os resultados de síntese para a arquitetura com CG, junto com um trabalho relacionado para a comparação. O trabalho relacionado (CONCEIÇÃO, 2014) apresenta uma IDCT para vários tamanhos de bloco (32x32 até 4x4). Destaca-se que este trabalho relacionado não implementa *clock-gating* e nem apresenta processamento fixo de 32 amostras por ciclo.

Como pode ser observado o trabalho desenvolvido teve um consumo maior de área que o trabalho (CONCEIÇÃO, 2014). Entretanto, ao se comparar resultados de dissipação de potência, arquitetura aqui apresentada obteve melhor

desempenhos nos dois casos comparados com (CONCEIÇÃO, 2014). No pior caso do trabalho relacionado, obtivemos uma redução de potência de 91,1%, e no melhor caso dele, ainda assim obtivemos uma redução de 6,2%. Salienta-se que a arquitetura desenvolvida utilizou uma frequência duas vezes maior que a utilizada pelo melhor caso do trabalho relacionado (CONCEIÇÃO, 2014).

Tabela.2 – Comparação com trabalhos relacionados.

Arquitetura	Área (K gates)	Tecnologia (nm)	Amostras por ciclo	Frequência (MHz)	Power total (mW)
Desenvolvida	356,2	45	32	93,3	30,02
(CONCEIÇÃO ,2014) pior caso	238,2	90	4	370	335,20
(CONCEIÇÃO ,2014) melhor caso	238,2	90	32	46,64	32,00

4. CONCLUSÕES

Este trabalho apresenta uma solução arquitetural para a IDCT do padrão de codificação de vídeo HEVC. A arquitetura desenvolvida é capaz de processar 32 amostras por ciclo independentemente do tamanho de TU a ser processada.

A utilização da técnica de *clock-gating* mostrou ser muito eficiente para esta aplicação, reduzindo em 66,2% a dissipação de potência do hardware. A arquitetura desenvolvida atingiu o processamento de 60 quadros por segundo para vídeos UHD 8K, dissipando 30,02 mW. Como trabalhos futuros seria realizar esta analise da arquitetura utilizando entradas reais, o que irão proporcionar resultados mais realistas para a arquitetura pelo fato de utilizar um cenário real de operação.

5. REFERÊNCIAS BIBLIOGRÁFICAS

AGOSTINI, L**. Desenvolvimento de Arquiteturas de Alto Desempenho Dedicadas a Compressão de Vídeo Segundo o Padrão H.264/AVC**. 2007. 172f. Tese (Doutorado em Ciência da Computação) – Instituto de Informática, UFRGS, Porto Alegre.

BOSSEN, F.; Bross, B.; Sühring, K., and D. Flynn, “HEVC complexity and implementation analysis**,” IEEE Trans. Circuits Syst. Video Technol**., vol. 22, no. 12, pp. 1684–1695, Dec. 2012.

GHANBARI, M. Standard Codecs: Image Compression to Advanced Video Coding. United Kingdom**. The Institution of Electrical Engineers**, 2003.

ITU, International Telecommunication Union. **ITU-T Recommendation H.264/AVC: Advanced Video Coding for generic audiovisual services**. Janeiro 2012. Disponivel em: http://handle.itu.int/11.1002/1000/11466.

JCT-VC, **ITU-T Recommendation H.265 - High Efficiency Video Coding** (ITU-T Rec.H.265), Abril 2013. Disponivel em: http://handle.itu.int/11.1002/1000/11885.

Pal, Ajit. **Low-Power VLSI Circuits and Systems**. Springer, 2014.

R. Conceição, J. C. de Souza, R. Jeske, M. Porto, B. Zatt and L. Agostini, "Power efficient and high troughtput multi-size IDCT targeting UHD HEVC decoders," **2014 IEEE International Symposium on Circuits and Systems (ISCAS)**, Melbourne VIC, 2014, pp. 1925-1928.

SULLIVAN, G. J.; Ohm, J.-R.; Han, W.-J.; Wiegand, T. . “Overview of the High Efficiency Video Coding (HEVC) standard,” **IEEE Trans. Circuits Syst. Video Technol.**, vol. 22, no. 12, pp. 1648–1667, Dec. 2012.