ARQUITETURA DE HARDWARE COM ELEVADA TAXA DE PROCESSAMENTO PARA O MODO PLANAR DE CODIFICAÇÃO DE VÍDEO DO PADRÃO HEVC

LUAN PIZZAMIGLIO AUDIBERT[1]; VLADIMIR AFONSO[2]; MARCELO PORTO[1]; BRUNO ZATT[1]; LUCIANO AGOSTINI[1]

[1]Universidade Federal de Pelotas – Grupo de Arquiteturas e Circuitos Integrados
[2]Instituto Federal Sul-rio-grandense
{lpaudibert[1], vafonso[2], porto[1], zatt[1], agostini[1]}@inf.ufpel.edu.br

1. INTRODUÇÃO

Aplicações e dispositivos capazes de manipular e reproduzir vídeos digitais em alta e ultra-alta resoluções estão surgindo a cada dia, como *smartphones*, *tablets* e TV Digital. Estes vídeos possuem uma grande quantidade de dados que precisam ser processados, armazenados e transmitidos, de forma que soluções eficientes para sua compressão são necessárias. Atualmente, o padrão estado-da-arte para a compressão de vídeos digitais é o *High Efficiency Video Coding* (HEVC) (ITU-T, 2015), o qual reduz em até 50% o *bit rate* quando comparado ao seu antecessor H.264/AVC (*Advanced Video Coding*) (ITU-T, 2003).

No entanto, soluções em software para codificadores de vídeo são impraticáveis para determinadas aplicações, seja pela dificuldade de atingirem processamento em tempo real ou por consumirem muita energia, inviabilizando a compressão em dispositivos alimentados por bateria. Portanto, hardwares dedicados que apresentem baixo consumo energético são imprescindíveis para este tipo de aplicação.

Um codificador baseado no padrão HEVC possui muitas etapas para a codificação de um vídeo, onde cada uma destas explora um determinado tipo de redundância. A etapa explorada neste trabalho é a Predição Intra-Quadro (ou apenas Intra). A Predição Intra é responsável por explorar a redundância espacial, ou seja, explora as similaridades das informações existentes dentro de um quadro, como regiões homogêneas e bordas de objetos estáticos.

No padrão HEVC, 35 modos da Predição Intra são permitidos e podem ser classificados em duas categorias: a) Modos de Predição Angular (33 modos), os quais permitem ao codificador modelar precisamente as estruturas com bordas direcionais; b) Modos de Predição Planar e DC, os quais promovem estimativas para regiões homogêneas (SZE, 2014).

O modo de predição Planar foi o escolhido para o desenvolvimento deste trabalho por se tratar do modo Intra mais utilizado no HEVC. De acordo com Corrêa (2015), o modo Planar é escolhido em 18,33% dos casos por trazer o melhor resultado de compressão. Neste trabalho, uma arquitetura com elevada taxa de processamento para o modo Planar do padrão HEVC é apresentada. A arquitetura considera blocos de tamanho 8x8 e é capaz de processar vídeos 8K (8192 x 4320 pixels) em tempo real.

2. METODOLOGIA

O modo Planar utiliza dois filtros para ser calculado, denominados de Horizontal e Vertical. Estes dois filtros são calculados através de multiplicações utilizando amostras de referência pertencentes aos blocos de imagem localizados acima e a esquerda do bloco que está sendo codificado. Os filtros de Predição

horizontal (Ph) e vertical (Pv) utilizam as equações (1) e (2) para serem calculados, respectivamente.

O valor de *N* nas equações representa o tamanho do bloco a ser predito, sendo que *x* e *y* ∈ *{0, …, N-1}*, enquanto que *p[-1][y]* e *p[-1][N]* são as amostras de referência localizadas a esquerda do bloco, *p[x][-1]* e *p[N][-1]* são as amostras localizadas acima deste bloco e *Ph[x][y]* e *Pv[x][y]* são as amostras preditas pelos filtros.

De posse dos resultados dos filtros horizontais e verticais, o modo Planar é obtido pela média destes resultados como mostrado pela equação (3), onde *p[x][y]* é o resultado final do modo Planar.

$$Ph[x][y] = (N-1-x) * p[-1][y] + (x+1) * p[N][-1] \quad (1)$$

$$Pv[x][y] = (N-1-y) * p[x][-1] + (x+1) * p[-1][N] \quad (2)$$

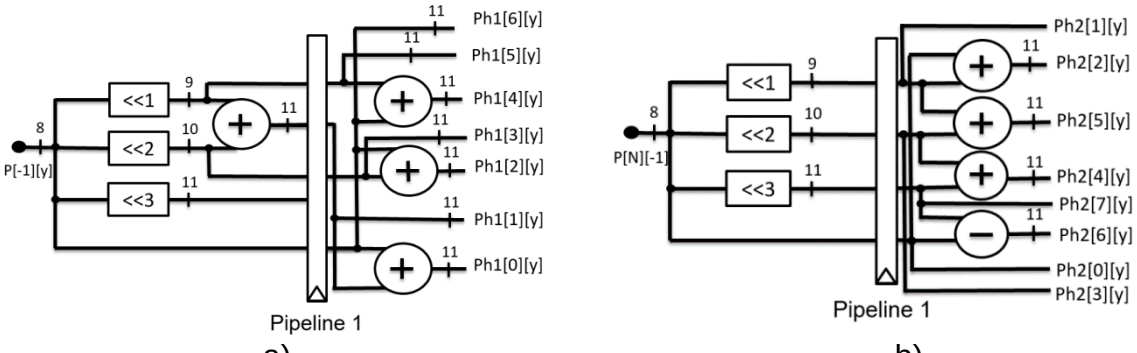
$$p[x][y] = (Ph[x][y] + Pv[x][y]) \gg (log_2(N)+1) \quad (3)$$

Baseado nas equações (1-3), uma arquitetura de hardware que implementa o modo Planar foi desenvolvida. Como os filtros Horizontal e Vertical utilizam as mesmas operações aritméticas, o mesmo projeto de hardware pode ser aproveitado para os dois filtros, mudando apenas as entradas utilizadas. Ou seja, para o cálculo das amostras do filtro Horizontal, o *pipeline* é preenchido com as amostras de referência à esquerda do bloco, e para o cálculo das amostras do filtro Vertical, o *pipeline* é preenchido com as amostras de referência acima do bloco que está sendo codificado.

Visto que multiplicadores são muito custosos em hardware, a arquitetura foi otimizada com a substituição de multiplicadores convencionais por operações de soma e deslocamento. Para que isto fosse possível, a arquitetura foi dividida em duas partes, como pode ser observado na Figura 1.

A primeira parte da arquitetura é mostrada na Figura1-a e realiza o cálculo *(N – 1 – x) * p[1][y]*. Como o valor de *N* é o tamanho do bloco a ser calculado, este pode ser fixado para o cálculo dos possíveis valores que irão multiplicar *p[-1][y]*. A entrada *p[-1][y]* é a amostra de referência que será usada no cálculo e as saídas *Ph1[0]-Ph1[7]* representam o resultado para uma linha ou coluna, onde *Ph1[7]* é sempre zero. Ainda, uma barreira de *pipeline* foi utilizada para que o caminho crítico tenha apenas um somador, fazendo com que a frequência máxima da arquitetura aumente.

A segunda parte da arquitetura é mostrada na Figura1-b e é responsável por calcular *(x+1) * p[N][-1]*. Esta metade da arquitetura também utilizou uma barreira de *pipeline*, visando o aumento da frequência máxima e também a sincronização das duas partes.

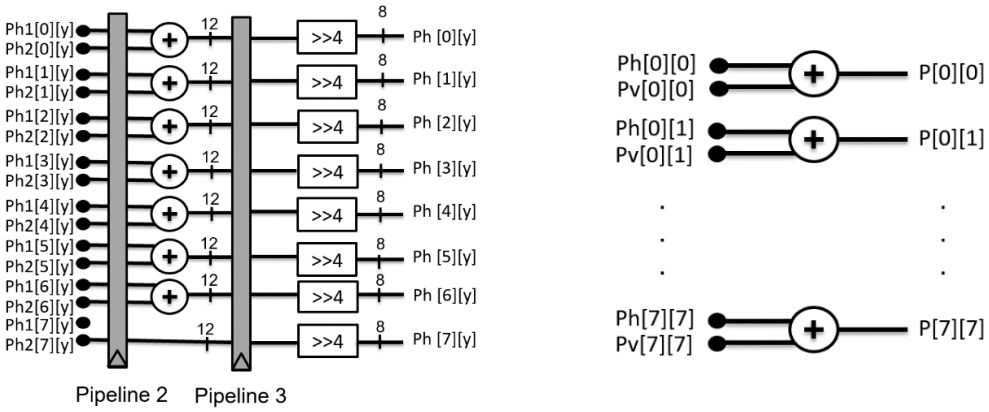


a) b)

Figura 1 – Arquitetura desenvolvida para os filtros Horizontal e Vertical dividida em duas partes.

Após o cálculo das duas partes do filtro Horizontal ou Vertical, é necessário o cálculo da média. A Figura 2-a mostra como as duas partes da arquitetura do filtro são unidas (*Ph1* e *Ph2*) para formar a Predição horizontal ou vertical completa. É importante ressaltar que *Ph1[7]* é sempre zero e, por isso, não tem a necessidade de somá-lo com *Ph2[7]* para formar o resultado. Após a soma, os resultados intermediários passam por um estágio de *pipeline* e, por fim, são deslocados quatro bits à direita para compor a média. Desta forma, uma linha ou coluna (no caso do filtro vertical) é obtida. Com o objetivo de obter o maior *throughput* possível, arquiteturas que calculam oito linhas e oito colunas do modo Planar (considerando um bloco 8x8) foram totalmente paralelizadas.

De posse dos resultados das predições horizontal e vertical para cada uma das amostras do bloco predito, estes dois valores devem ser somados para obtenção do modo Planar Completo. Neste trabalho, optou-se por utilizar um somador para cada amostra a fim de obter o maior *throughput* possível. A Figura 2-b mostra a arquitetura responsável por esta etapa final, onde 64 somadores são utilizados para compor o resultado final do modo Planar.



a) b)

Figura 2 – Etapa final da arquitetura proposta.

3. RESULTADOS E DISCUSSÃO

Para comparação, três versões diferentes da arquitetura desenvolvida foram implementadas e sintetizadas utilizando o software Altera Quartus II na versão 13.0sp2 (Altera, 2013) e o dispositivo Stratix V 5SGXEA5N1F45C1. Porém, para mostrar os reais ganhos das otimizações, uma versão com somente as otimizações da ferramenta da Altera foi gerada, a qual foi chamada de *Versão da Ferramenta*. A arquitetura descrita com todas as três barreiras de *pipeline*, foi chamada de *4 Estágios de Pipeline*. A terceira versão considera apenas o pipeline 2 ativo,e foi chamada de *2 Estágios de Pipeline*. Por fim, uma quarta versão da arquitetura descrita, sem a utilização de pipelines, é denominada de *Versão Combinacional*. Os resultados destas arquiteturas são apresentados na Tabela 1.

Como pode ser observado na Tabela 1, a *Versão da Ferramenta* foi a que apresentou maior utilização de recursos de hardware e menor frequência máxima de operação entre todas as soluções, mesmo quando comparada a versão *4 Estágios de Pipeline*. A versão *4 Estágios de Pipeline*, como esperado, foi a que atingiu maior frequência de operação e utilizou mais recursos de hardware dentre as arquiteturas desenvolvidas. É possível observar também que a versão *2 Estágios de Pipeline* pode operar com uma frequência maior do que a *Versão Combinacional*, porém utilizando 1984 registradores. A *Versão Combinacional* da arquitetura se mostrou suficiente para o processamento de vídeos 8K (8192 x

4320 pixels) em tempo real com o menor custo em termos de hardware e menor frequência de operação dentre as versões avaliadas. Apesar de existirem outras arquiteturas para o cálculo do modo Planar na literatura, estes trabalhos não permitem uma comparação justa com a arquitetura desenvolvida, pois apresentam seus resultados considerando vários tamanhos de bloco do modo Planar e não apenas o tamanho 8x8, foco deste trabalho.

Tabela 1 – Resultados da síntese em FPGA para o modo Planar Completo.

Versão da arquitetura	Versão da Ferramenta	4 Estágios de Pipeline	2 Estágios de Pipeline	*Versão Combinacional*
Área (ALMs)	6771	1872	1038	899
Registradores	0	4016	1984	0
Frequência Máxima (MHz)	169	827	512	380
Ciclos p/ processar um bloco 8x8	1	4	2	1
Freq. para resolução 4320p@120qps (MHz)	66,36	265,42	132,71	66,36

4. CONCLUSÕES

Este trabalho apresentou diferentes versões de uma arquitetura de alto desempenho para o modo Planar de Predição Intra do padrão HEVC capaz de processar blocos de tamanho 8x8. Com os resultados obtidos, a arquitetura desenvolvida é capaz de processar vídeos em resolução 8K (8192 x 4320 pixels) em tempo real a uma taxa de 120 quadros por segundo operando a uma frequência de apenas 66,36 MHz na sua *Versão Combinacional*. É importante mencionar que a *Versão Combinacional* da arquitetura desenvolvida reduz em aproximadamente 87% a utilização de recursos de hardware quando comparada com a versão otimizada apenas pela ferramenta Altera Quartus II, o que resulta em grande redução no consumo de energia. Como trabalhos futuros pretende-se desenvolver uma arquitetura de hardware capaz de processar todos os tamanhos de bloco em que o modo Planar pode ser utilizado no padrão HEVC e gerar resultados de síntese considerando tecnologia ASIC e estimativas de dissipação de potência, tanto para a arquitetura desenvolvida, como para as próximas arquiteturas a serem implementadas.

5. REFERÊNCIAS BIBLIOGRÁFICAS

ALTERA. **Quartus II Web Edition**. Acessado em: 09 jul. 2016. Online. Disponível em: http://dl.altera.com/13.0sp1/?edition=web

CORRÊA, M; PORTO, M; ZATT, B; AGOSTINI, L. A Low-Area and High-Throughput Intra Prediction Architecture for a Multi-Standard HEVC and H.264/AVC Video Encoder. **Symposium on Integrated Circuits and System Design**. Salvador, v28, 2015.

ITU-T. International Telecomunication Union. **ITU-T Recommendation and Final Draft International Standard of Joint Video Specification**. Maio, 2003.

ITU-T. International Telecomunication Union. **Recommendation ITU-T H.265: High Efficiency Video Coding**. Abril, 2015.

LAINEMA, J; WOO-JIN, H; Intra-Picture Prediction in HEVC. SZE, V. **High Efficiency Video Coding (HEVC) Algorithms and Architectures**. New York: Springer, 2014. Cap.4, p.91-112.