Redes neurais para filtragem inversa com aplicação em calorímetros operando a alta taxa de eventos

Mateus H. M. Faria, Luciano M. A. Filho, João Paulo B. da S. Duarte e José M. Seixas

***Resumo*— Este trabalho tem como objetivo a implementação de um sistema inverso para remover a resposta ao impulso de um calorímetro operando a uma alta taxa de eventos. O calorímetro Tilecal apresenta uma resposta ao impulso maior do que a taxa de eventos adquiridos, gerando empilhamento de sinais na sua eletrônica de leitura. Foram propostos filtros FIR que aproximam o sistema inverso, no entanto, características não-lineares intrínsecas ao calorímetro como desvio de fase e deformação do pulso, fazem com que filtros lineares não sejam adequados a este cenário. Devido à essas características, é proposto um filtro inverso baseado em redes neurais, implementável em hardware dedicado (FPGA). Este método mostrou-se mais eficiente, ao ser comparado à um método de filtragem baseado em filtro linear.**

***Palavras-Chave*— Rede neural, FPGA, calorímetro, empilhamento de sinais.**

***Abstract*— This paper aims to implement an inverse system to remove a impulse response from a calorimeter operating at a high event rate. The Tilecal calorimeter has a higher impulse response than the acquired event rate, generating signal pile-up in its reading electronics. FIR filters have been proposed that approximate the inverse system, however, nonlinear characteristics intrinsic to the calorimeter such as phase shift and pulse deformation, make linear filters not suitable for this scenario. Due to these characteristics, an inverse filter based on neural networks, implemented in dedicated hardware (FPGA), is proposed. This method proved to be more efficient when compared to filtering method based on linear filter.**

***Keywords*— Neural network, FPGA, calorimeter, signal pile-up.**

I. Introdução

Os sistemas de instrumentação se desenvolvem a partir da necessidade, imposta, muitas vezes, por requisitos altamente específicos de aplicações industriais e experimentos científicos. Os experimentos em física de altas energias, que estudam os constituintes básicos da matéria e as forças de interação entre eles, fazem o uso de um complexo sistema de instrumentação que, devido à raridade dos eventos de interesse, deve atender aos requisitos de precisão e exatidão das variáveis em análise funcionando à uma alta taxa de eventos.

O Grande Colisor de Hádrons (LHC), é o maior acelerador de partículas já construído, tendo entrado em operação em setembro de 2008. Ele opera acelerando feixes de prótons em um túnel subterrâneo de 27 km de circunferência [1], [2], incidindo-os uns sobre os outros no ponto de interesse de quatro detectores (experimentos) principais: ALICE [3], LHCb [4], CMS [5] e ATLAS [6].

Mateus H. M. Faria†, Luciano M. A. Filho†, João Paulo B. da S. Duarte†, José M. Seixas‡, Unversidade Federal de Juiz de Fora (UFJF)†, Juiz de Fora-MG, Brasil, Universidade Federal do Rio de Janeiro (UFRJ)‡, Rio de Janeiro-RJ, Brasil. E-mails: mateus.hufnagel@engenharia.ufjf.br, luciano.andrade@engenharia.ufjf.br, joao.duarte@engenharia.ufjf.br, seixas@lps.ufrj.br

A. O experimento ATLAS

Quando uma colisão altamente energética ocorre em um acelerador de partículas, ocorrem fenômenos como a produção espontânea de outras partículas elementares. Para identificar as partículas que foram geradas, utilizam-se os produtos provenientes de decaimentos das partículas. Para isso, os detectores são projetados em camadas, visando identificar e medir propriedades e grandezas de um evento como massa, velocidade e carga das partículas.

O ATLAS é o maior dentre os experimentos do LHC, sendo ilustrado na Figura 1. O ATLAS possui formato cilíndrico, sendo composto de três sub-sistemas: detector de trajetória (mais interno), calorímetros eletromagnético e hadrônico e, na camada mais externa, o espectrômetro de múons [6].

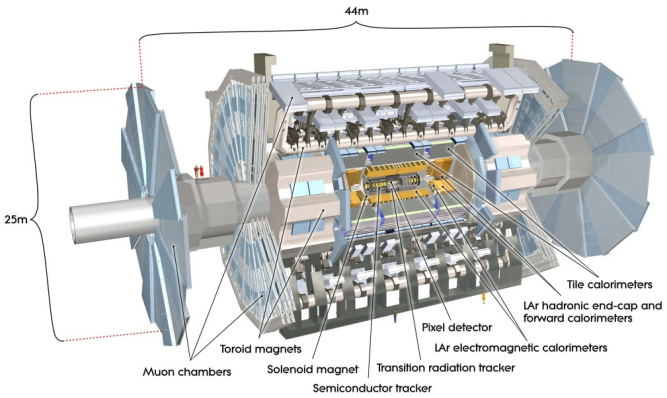

Fig. 1. Visão geral dos experimentos do LHC. Extraído de [6].

Os calorímetros são instrumentos que têm como funcionalidade a exploração dos chuveiros de partículas produzidos pelo material detector, com o intuito de medir a energia das partículas produzidas nas colisões [7]. Ocupam geralmente o maior volume dentre os detectores de partículas. O calorímetro hadrônico, ou de telhas (Tilecal), é um calorímetro por amostragem, assim chamado por constituir-se de chapas de aço como material absorvedor, intercaladas com telhas de material cintilante, que emitem luz ao interagirem com as partículas incidentes. É um sistema que apresenta uma fina segmentação, totalizando cerca de 10.000 canais de leitura, o que permite uma boa localização espacial do chuveiro de partículas [6], [7].

O chuveiro de partículas, ao interagir com uma célula, produz um sinal luminoso que é convertido em sinal elétrico pelas PMT's. Este sinal é condicionado, amplificado e digitalizado, possuindo uma amplitude proporcional à energia depositada na célula. A incidência de uma partícula proveniente dos eventos de colisão gera, nos canais de leitura de cada célula, um pulso

com uma duração total de cerca de 150 ns, como ilustrado na Figura 2 [8].

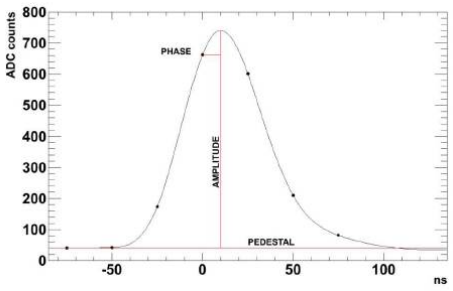

Fig. 2. Pulso analógico típico do Tilecal digitalizado em um conjunto de 7 amostras. A taxa de amostragem do ADC é sincronizada com o cruzamento de feixe do LHC. Extraído de [8].

Atualmente, o LHC trabalha colidindo prótons a uma taxa de 40 milhões por segundo, ou seja, os feixes se cruzam no ponto de interesse dos detectores a cada 25 ns. Como os eventos de interesse são raros, nas atualizações futuras, a luminosidade do LHC irá aumentar [9]. Este parâmetro é fundamental em aceleradores de partículas, visto que quanto maior, mais interações ocorrem, em média, por cruzamento [10].

Com o aumento da probabilidade de interação próton-próton em cada cruzamento de feixes, haverá uma maior produção de eventos provenientes das colisões. Isto acarretará em uma maior chance de deposições sucessivas de energia, em uma mesma célula, em cruzamentos de feixes subsequentes. Logo, o pulso característico do canal de leitura das células sujeitas a estas condições pode deformar, ocasionando erros de medição da energia depositada na mesma. A este efeito, dá-se o nome de empilhamento de sinais, ou *pile-up*, sendo ilustrado na Figura 3. Deste modo, neste trabalho, será utilizado o termo *ocupância* para quantificar células que apresentam intensidades de empilhamento de sinais distintas. Ou seja, quanto maior a ocupância, maior a probabilidade de incidência de partículas em uma mesma célula, em cruzamentos de feixe seguidos. Dá-se o valor de zero à cem porcento para essa chance de deposição.

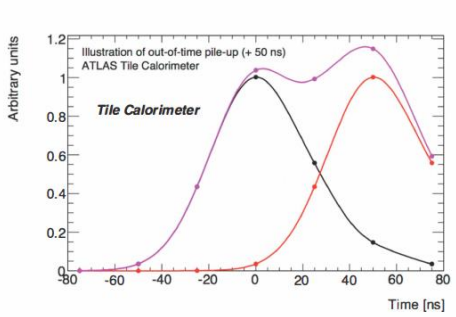

Fig. 3. Ilustração do efeito de empilhamento de sinais (*pile-up*). Extraído de [11].

Após passar pelas atualizações, que estão planejadas para acontecer em 2024, devido a alta resolução, tanto do Tilecal, quanto dos demais detectores, todo o experimento ATLAS irá gerar uma quantidade de dados ainda maior a cada cruzamento de feixe, aumentando ainda mais a incidência de empilhamento de sinais. Para lidar com estas limitações, foram propostas técnicas de filtragem inversa utilizando filtros FIR, em que simulações demonstraram uma relativa melhoria no desempenho desta estimação [14]. No entanto, distúrbios de natureza não-linear, intrínsecos dos canais de leitura do Tilecal, como ruído de fase proveniente do tempo de voo das partículas e ruído eletrônico, não são modelados de maneira ótima por métodos lineares.

II. Método proposto

Este trabalho propõe uma solução de pré-processamento para estimação de energia dos sinais de leitura das células do Tilecal em uma FPGA, que realizará uma filtragem inversa não-linear, com o intuito de recuperar apenas a amostra de energia no cruzamento de feixe referente à sua deposição real. Para a resolução deste problema, parte-se da interpretação do sinal de leitura do calorímetro como o resultado da convolução linear da amostra de energia teórica armazenada na célula, adicionando os efeitos de tempo de vôo das partículas e ruído da eletrônica de leitura.

A obtenção de um modelo matemático para realizar a compensação da influência de um canal sobre o sinal transmitido, principalmente quando o mesmo possui características não-determinísticas e não-linearidades, pode se tornar muito complexo. Sendo assim, são aplicadas técnicas de Least Squares (LS) [14] para estimação de um sistema inverso do modelo desconhecido, porém, atua fora das condições ótimas. Como alternativa, topologias distintas de redes neurais artificiais como RBF, redes recursivas, FLANN e MLP são largamente aplicadas, buscando a modelagem das características não-lineares de canais, muito utilizadas, por exemplo, em telecomunicações [15].

Como o número de canais de leitura do calorímetro no cenário de alta luminosidade (pós-atualização) é alto, além de bom desempenho, o sistema de processamento deve possuir baixa latência e ser compacto o suficiente, permitindo sua implementação em um hardware dedicado. De acordo com estes requisitos, deve-se escolher parâmetros como a topologia, número de neurônios, nós de entrada, o número de pesos e a precisão requerida para representá-los de acordo com o critério de desempenho estabelecido. Com o intuito de simplificação, o ajuste dos pesos sinápticos (treinamento) será realizado em software [16].

A topologia escolhida será a de uma rede MLP, visto que não possui realimentação e por possuir um alto grau de paralelismo, permite uma implementação com baixa latência em FPGAs. Além disso, funções de ativação não-lineares a torna adequada para as condições do canal. O número de neurônios e nós de entrada serão definidos através simulações exaustivas, buscando uma rede que possua a melhor relação entre o número de pesos sinápticos e o desempenho da mesma.

A arquitetura utilizada neste trabalho consiste em uma rede perceptron com uma camada escondida, contendo múltiplos neurônios com função de ativação tangente hiperbólica e combinação linear na camada de saída, com apenas um neurônio. Os nós de entrada da rede representam uma janela

do sinal do canal de leitura, armazenados em memórias do tipo *pipeline*. O ajuste dos pesos sinápticos da rede foi realizado pelo algoritmo de treinamento supervisionado Levenberg-Marquardt, utilizando como critérios de parada a checagem por validação cruzada e o número de épocas, evitando o *overtraining*. Cada um destes treinamentos será inicializado 30 vezes, sendo escolhida a rede com o melhor desempenho. Este, é dado pela raiz do erro médio quadrático (RMSE) entre o valor estimado e o valor teórico da energia. Todos os sinais utilizados para treinamento, validação e teste da rede foram obtidos através simulações de *Toy* Monte Carlo, que gera características semelhantes às do Tilecal [17].

Foram gerados nove conjuntos de sinais, contendo os sub-conjuntos abaixo:

- Treinamento/Validação – 50.000 amostras;
- Teste – 100.000 amostras;

Cada conjunto é composto de sinais de células com um valor de ocupância de sinal distinto, de 10 % a 90 %, com o intuito de realizar os ajustes dos pesos sinápticos para diferentes intensidades de ocorrência do efeito de empilhamento de sinais. Logo, cada sub-conjunto de treinamento é composto de pares de sinal de entrada/saída desejada $\{\mathbf{y}, x\}$, contendo amostras de sinais de células com ocupâncias distintas. Os sinais de entrada representam a leitura de um canal de uma célula do calorímetro, sendo o número de amostras no vetor $\mathbf{y}$ igual ao número de nós de entrada da rede $(m+1)$. A saída desejada x consiste de uma amostra da amplitude do sinal sem as distorções do Tilecal, referente à amostra central à janela, $\frac{m}{2}$, no caso da rede possuir número de nós par, e $\frac{m+1}{2}$ para números ímpares. Após o ajuste adequado dos pesos sinápticos, a rede neural deve fornecer uma saída que é a estimativa da amplitude da energia da amostra ao centro ou próxima dele, para as possíveis arquiteturas de entrada, exemplificado na Figura 4.

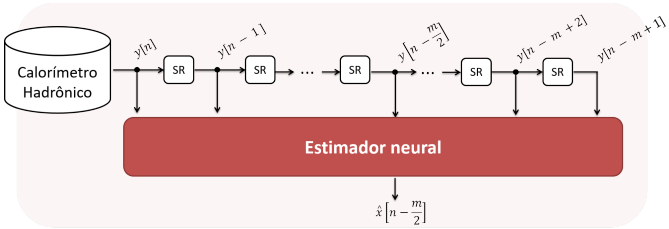

Fig. 4. Rede neural *feedforward* para aplicação *online* com janela par.

O melhor estimador neural encontrado terá seus pesos sinápticos quantizados utilizando a aritmética de ponto fixo, sendo que o número de bits necessários para representar cada constante será otimizado visando melhor desempenho (baixa latência) e menor ocupação da área em uma FPGA. O critério de erro adotado entre a rede com os pesos idealmente representados em software e os mesmos quantizados, foi uma tolerância do erro relativo de até 1 %. As funções de ativação serão representadas em hardware através de *Look-up tables* (LUTs).

III. Resultados

O ajuste do número de neurônios da camada oculta e do numero de nós de entrada da rede foi realizado de forma conjunta, pois apresentam forte interdependência. Nas Figuras 5 e 6 é possível observar superfícies, geradas à partir do menor valor de erro obtido para cada combinação entre número de nós de entrada e neurônios na camada oculta para um sinal de teste com 10 % e 90 % de ocupância, respectivamente. Notou-se que para sinais de células com baixa ocupância, atingiu-se um bom desempenho com 5 atrasos e 4 neurônios na camada escondida. Ao testar o estimador neural com sinais de ocupâncias mais altas, o número de neurônios mostrou-se suficiente. No entanto, para atingir um melhor desempenho, necessitou-se de mais nós de entrada, chegando a 10, para sinais de células com 90 % de ocupância. Logo, a rede escolhida para atuar em todos os cenários é uma rede 10-4-1.

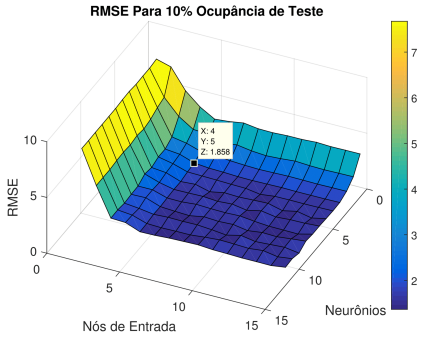

Fig. 5. Erro RMS para redes com números de neurônio e nós de entrada distintos, para um sinal de teste com 10 % de ocupância.

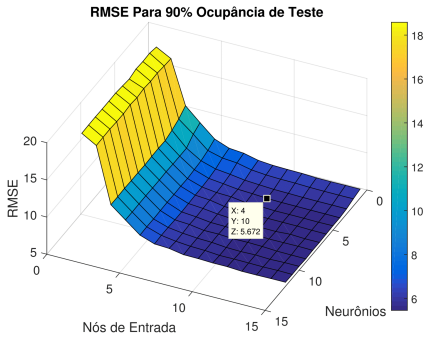

Fig. 6. Erro RMS para redes com números de neurônio e nós de entrada distintos, para um sinal de teste com 10 % de ocupância.

Na Figura 7, é possível observar uma comparação do desempenho entre o método proposto e o baseado em filtros FIR, onde o estimador neural apresentou um melhor desempenho para células com qualquer valor de ocupância.

A. *Interpretação dos pesos da rede*

Realizar uma análise de baixo nível em uma rede neural pode trazer uma nova perspectiva sobre um problema, proporcionando interpretações de como os dados são processados por suas unidades. Seguindo esta linha de pensamento, a rede

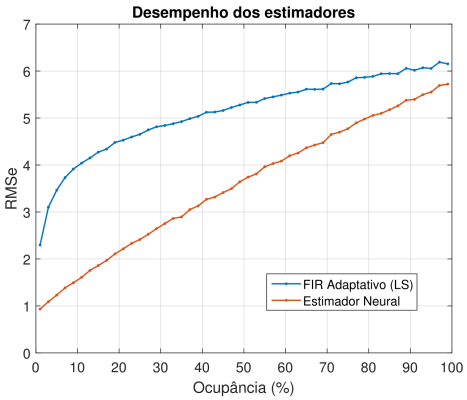

Fig. 7. Comparação do desempenho entre o estimador neural e o filtro FIR, baseado em Least Squares.

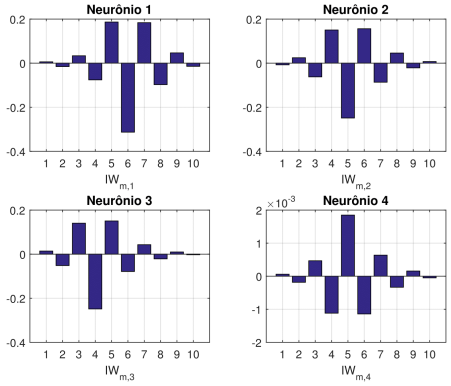

Fig. 9. Pesos da camada de entrada (sem bias).

10-4-1, após ter seus pesos ajustados pelos processos descritos nas seções anteriores, foi submetida à uma análise baseada em histogramas bidimensionais (*scatter plot*), conforme ilustrada na Figura 8.

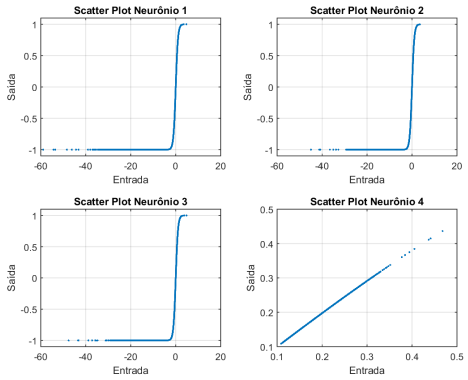

Fig. 8. *Scatter plot* dos neurônios (entrada e saída da tangente hiperbólica) da camada escondida para sinal de teste com 30 % de ocupância.

Nota-se que, os neurônios 1, 2 e 3, têm amostras ativando uma grande faixa de saturação negativa da tangente hiperbólica e, também, no domínio o qual a função possui maior derivada, como normalmente espera-se ao utilizar funções de ativação não-lineares. No entanto, o neurônio 4 teve apenas uma faixa aproximadamente linear da tangente hiperbólica ativada. Esses resultados estão intimamente conectados aos pesos da camada de entrada para cada um dos neurônios, ilustrados na Figura 8.

Os pesos do neurônio 4 possuem uma configuração diferente dos demais, além de sua escala de valores estar confinada na ordem de 10^{-3}. A amplitude dos pesos revela que este neurônio em particular, está modificando a amplitude do sinal de entrada, confinando-o no domínio de derivada aproximadamente constante da tangente hiperbólica. Uma hipótese para essa constatação, é que este neurônio pode ter sido configurado no treinamento para lidar com as características intrinsecamente lineares do sinal em questão, aproximando-se de um filtro linear FIR de ordem 9, como é comparado na Figura 10. Salvo uma constante de escala entre os dois conjuntos de valores, suas estruturas são bastante similares, reforçando a hipótese de que a rede neural, durante o treinamento, encontrou uma solução ao qual realiza uma filtragem linear através de um neurônio, com uma correção não-linear, realizada pelos demais em paralelo.

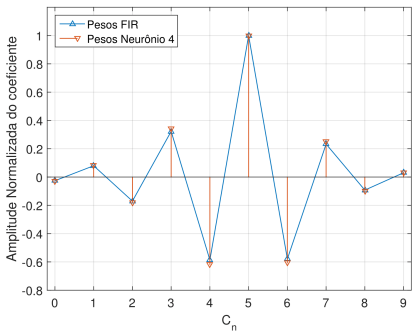

Fig. 10. Pesos de entrada do neurônio 4 comparados aos de um filtro FIR de ordem 9.

Outro elemento reforçador para esta hipótese são os pesos da camada escondida, mostrados na Tabela I, novamente, destacando o papel do neurônio 4. O treinamento compensou os valores de saída pequenos do neurônio, devido a sua escala dos pesos de entrada, com um peso muito grande nesta camada, intensificando a ação que este neurônio tem na resposta final da rede neural, aproximando, em escala, dos pesos do filtro FIR citado anteriormente.

TABELA I

PESOS SINÁPTICOS NA CAMADA ESCONDIDA.

Neurônio	1	2	3	4
Peso sináptico	-5,05	10,25	-7,48	1.358,23

Ao observar o histograma do sinal de cada neurônio, antes de serem combinados pelo neurônio linear de saída com bias, ilustrado na Figura 11, é possível inferir que a maior parte da informação de saída da rede vem do neurônio com ativação linear, enquanto que os outros três neurônios aplicam somente um ajuste fino não-linear, em termos de valores de ADC, à resposta final. Portanto, apesar de um filtro FIR com 10 coeficientes não realizar a estimação de forma muito eficiente, como demonstrado em [14], no qual foram utilizados 26

TABELA II
COMPARAÇÃO ENTRE AS ARQUITETURAS IMPLEMENTADAS EM FPGA.

Pipelines	Elementos lógicos	Fmáx (MHz)	Uso de memória (Kbits)	Latência em BC
0	13.319	15,07	0	5
1	1.413	34,24	491,52	6
2	1.419	41,93	491,52	7

coeficientes para atingir a melhor performance que o mesmo possa alcançar no ambiente em questão, ao utilizar neurônios em paralelo para correção não-linear, a rede neural consegue estimar a energia do sinal em uma janela de sinal mais estreita, favorável para trabalhar em um ambiente com alta taxa de eventos.

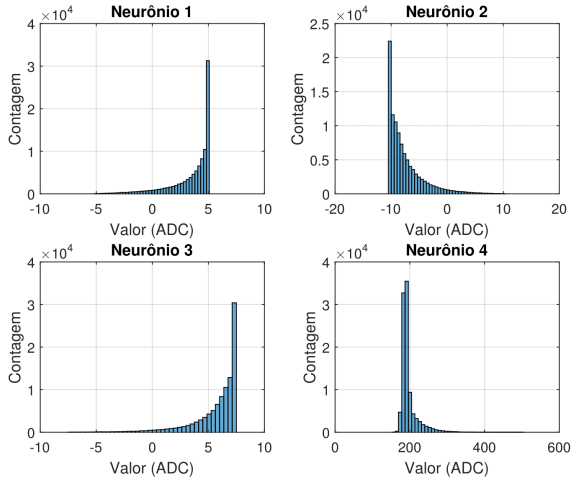

Fig. 11. Histograma da saída de cada neurônio antes de sua combinação linear com bias da camada de saída.

B. Implementação em FPGA

Inicialmente, foi testada uma rede totalmente combinacional. A frequência máxima de processamento alcançada foi de 16,78 MHz, muito abaixo do necessário para funcionar no ambiente desejado, sendo que a quantidade de recursos consumida torna essa implementação inviável. Visto isso, foi utilizada uma arquitetura em *pipeline*. A utilização deste recurso é explorada em soluções de processamento digital de sinais, podendo ser aplicado em FPGAs sem excessiva utilização de elementos lógicos adicionais, visto que utilizam apenas *flip-flops*. Esse recurso permite a realização de operações em paralelo, tornando partes do circuito sincronizadas. Ao fim, para obter o resultado do processamento, aumenta-se um ciclo de *clock* para cada *pipeline*. No entanto, esta técnica possibilita a implementação desse circuito à taxas mais altas [12]. São propostas duas arquiteturas com 1 e 2 *pipelines*, no qual os resultados das simulações são encontrados na Tabela II. Conclui-se, então, que é necessário um *pipeline* de 2 estágios para atingir a frequência de operação requerida, que é de 40 MHz.

IV. CONSIDERAÇÕES FINAIS

O cenário de aumento de luminosidade previsto no LHC afeta diretamente os algoritmos de estimação de energia do Tilecal. Este trabalho propôs o uso de uma rede neural como uma alternativa para estimação de energia. Ao fim das análises, pode-se concluir que o estimador neural apresentou desempenho superior quando comparado a um método linear, visto sua capacidade de representações de funções complexas, através dos neurônios com função de ativação não-linear. Para trabalhos futuros, observa-se que, com uma análise dos pesos da rede, podem ser propostas novas formas para realizar o treinamento visando a obtenção de um estimador implementável em hardware com a utilização de menos recursos computacionais.

AGRADECIMENTOS

Os autores gostariam de agradecer ao CNPq, CAPES, FAPERJ, FAPESB, RENAFAE, MCTI, CERN (Suíça) e União Europeia pelo apoio financeiro.

REFERÊNCIAS

[1] CERN, C. *How an accelerator works?* Acessado em Abril de 2017. Disponível em: `http://home.cern/about/how-accelerator-works`

[2] EVANS, L., BRYANT, P. *LHC machine*. Journal of Instrumentation, IOP Publishing, v. 3, n. 08, p. S08001, 2008.

[3] ALICE, C. *The alice experiment at the cern lhc*. Journal of Instrumentation, IOP Publishing, v. 3, n. 08, p. S08002, 2008.

[4] LHCB, C. *The lhcb detector at the lhc*. Journal of instrumentation, IOP Publishing, v. 3, n. 08, p. S08005, 2008.

[5] CMS, C. *The cms experiment at the cern lhc*. JInst, Citeseer, v. 3, n. 08, p. S08004, 2008.

[6] ATLAS, C. *The atlas experiment at the cern large hadron collider*. Journal of Instrumentation, v. 3, n. 08, p. S08003, 2008.

[7] Anderson, K. et al. *Design of the front-end analog electronics for the ATLAS tile calorimeter*. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, v. 551, n. 2, p. 469-476, 2005.

[8] Peralva, B. S. *The TileCal energy reconstruction for collision data using the matched filter*. In: Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC). IEEE, 2013. p. 1-6.

[9] Cerqueira, A. S. et al. *Tile Calorimeter Upgrade Program for the Luminosity Increasing at the LHC*. arXiv preprint arXiv:1509.08994, 2015.

[10] Herr, W., Muratori, B. *Concept of luminosity*. In: proceedings of CERN Accelerator School. 2003. v. 361.

[11] Clement, C.; Klimek, P. *Identification of pile-up using the quality factor of pulse shapes in the ATLAS tile calorimeter*. In: Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC). IEEE, 2011. p. 1188-1193.

[12] Meyer-Baese, U. *Digital Signal Processing with Field Programmable Gate Arrays*. 3rd. edition. Springer Publishing Company, Incorporated, 2007. ISBN 3540726128, 9783540726128.

[13] Nakahama, Y. *The atlas trigger system: Ready for run-2*. In: IOP PUBLISHING. Journal of Physics: Conference Series. 2015. v. 664, n. 8, p. 082037.

[14] Duarte, J. B. d. S. E*studo de técnicas de deconvolução para reconstrução de energia online no calorímetro hadrônico do ATLAS*. Dissertação, PPEE UFJF, 2015.

[15] Burse, K., Yadav, R. N., Shrivastava, S. *Channel equalization using neural networks: A review*. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), IEEE, v. 40, n. 3, p. 352–357, 2010.

[16] Jamal, M., Sayed A. I. *Hardware Aspects of Artificial Neural Network and Its Applications*. International Conference on Advances in Computer, Electronics and Electrical Engineering, 2012, p. 460–465. ISBN: 978-981-07-1847-3.

[17] Chapman, J. *ATLAS simulation computing performance and pile-up simulation in ATLAS*. In LPCC detector simulation workshop, CERN, 2011.