

Redes Neurais Embarcadas em FPGA para Classificação Online de Elétrons em um Calorímetro de Altas Energias

Igo Amauri Dos S. Luz^{1,2}, Gabriel H. B. Lisboa³, Nestor Pereira Neto², João Victor da Fonseca Pinto³, Leandro Assis dos Santos³, Fernanda D. V. R. Oliveira³, Natanael N. M. Junior³, Edmar Egídio P. de Souza², José Manoel de Seixas³, Paulo C. M. A. Farias², Eduardo F. Simas Filho², Denis Oliveira Damazio⁴

¹Instituto Federal da Bahia, IFBA/SAJ

²Laboratório de Sistemas Digitais, PPGEEC/UFBA

³Laboratório de Processamento de Sinais, COPPE/POLI, UFRJ

⁴Brookhaven National Laboratory, BNL-EUA

Emails: {igo.luz@ifba.edu.br}, {gabriel.lisboa, jodafons, leandro.santos, fernanda.oliveira, natmourajr, seixas}@lps.ufrj.br, {nestor.dias, edmar.egidio, eduardo.simas, paulo.farias}@ufba.br, denis.oliveira.damazio@cern.ch

Resumo—O *Large Hadron Collider* (LHC), tem passado por atualizações para aumentar a quantidade de eventos simultâneos e a energia das colisões próton-próton. Para a *Run-4*, a energia de colisão aumentará em uma ordem de magnitude, e o número médio de interações por colisão alcançará uma média de 200. Para lidar com este aumento na taxa de eventos, será feita uma atualização do sistema de *Trigger and Data Acquisition* (TDAQ) do experimento ATLAS, no qual o *Global Trigger* será adicionado ao novo TDAQ. Ele implementará a seleção de partículas em hardware digital - FPGA (*Field Programmable Gate Array*), com uma taxa de saída de eventos de 1 MHz. Neste trabalho são apresentados os resultados da implementação em hardware do algoritmo para seleção online de elétrons, chamado *NeuralRinger*. Este método captura a informação lateral e longitudinal dos chuveiros eletromagnéticos, alimentando redes neurais para selecionar elétrons de modo eficiente. Os resultados de erro obtido com o circuito responsável pela construção dos anéis é da ordem de 1,4% e com o circuito da rede neural é da ordem de 2%. Desta forma, os resultados da implementação do *NeuralRinger* para a FPGA indicam boa concordância do algoritmo implementado em hardware, se comparado aos resultados de referência obtidos a nível de software. Os resultados de síntese do circuito indicaram a viabilidade de embarcar o circuito no *framework* do *Global Trigger*, com 1,42% de utilização de *LookUp Tables* e 0,96% de *Flip-Flops* para o circuito de construção dos anéis, e 0,12% de *LookUp Tables* e 0,26% de *Flip-Flops* para a rede neural.

Palavras chave—ATLAS experiment, Trigger, Field Programmable Gate Array, electrons.

I. INTRODUÇÃO

Experimentos de Física de Partículas dedicam-se à investigação dos constituintes elementares da matéria e suas interações fundamentais [1]. Um instrumento crucial para o avanço experimental desta área é o *Large Hadron Collider* (LHC), localizado no CERN, que possibilita o estudo de colisões de prótons a energias da ordem de TeV¹. Experimentos conduzidos nesta instalação culminaram em avanços

significativos, notadamente a descoberta do bóson de Higgs em 2012, realizada pelos experimentos *A Toroidal LHC Apparatus* (ATLAS) [2] e *Compact Muon Solenoid* (CMS) [3]. Este feito não apenas corroborou previsões teóricas fundamentais, mas garantiu um prêmio Nobel aos pesquisadores responsáveis [4]. A exploração de fenômenos que transcendem o Modelo Padrão constitui, atualmente, uma diretriz primordial da pesquisa na área, impulsionada pela necessidade de explicar questões relacionadas às forças fundamentais da natureza e a composição de blocos básicos de matéria [5].

O LHC é o maior e mais potente acelerador de partículas do mundo, com 27 km de circunferência e situado a 100 metros de profundidade [5]. Em seu interior, feixes de prótons são impulsionados em direções opostas para colidir em quatro pontos de interação estrategicamente definidos [6]. Nestas localizações, experimentos de grande porte – *A Large Ion Collider Experiment* (ALICE) [7], *Large Hadron Collider beauty* (LHCb) [8], CMS [9] e ATLAS [10] – são responsáveis por registrar detalhadamente os produtos emergentes dessas colisões, fornecendo assim os dados experimentais vitais para a investigação de processos subatômicos [6].

O experimento ATLAS, caracterizado por sua geometria cilíndrica (conforme ilustrado na Figura 1), é o maior experimento do LHC [5]. Foi projetado como um detector de propósito geral, capaz de estudar uma ampla variedade de processos físicos resultantes das colisões no LHC [11]. Sua estrutura é composta por quatro subsistemas: o detector de trajetórias, responsável pela reconstrução das trajetórias de partículas carregadas [12]; os calorímetros eletromagnéticos e hadrônicos, destinados à medição da energia depositada pelas partículas [13]; e o espectrômetro de múons, especificamente concebido para a identificação e medição dos múons [14]. A disposição em camadas destes subsistemas permite uma reconstrução abrangente dos eventos de colisão.

A operação do ATLAS em condições de alta taxa de eventos apresenta desafios significativos. Com colisões ocorrendo a

¹O (eV) é uma unidade definida como a quantidade de energia cinética ganha por um elétron quando submetido a uma diferença de potencial de 1V no vácuo.

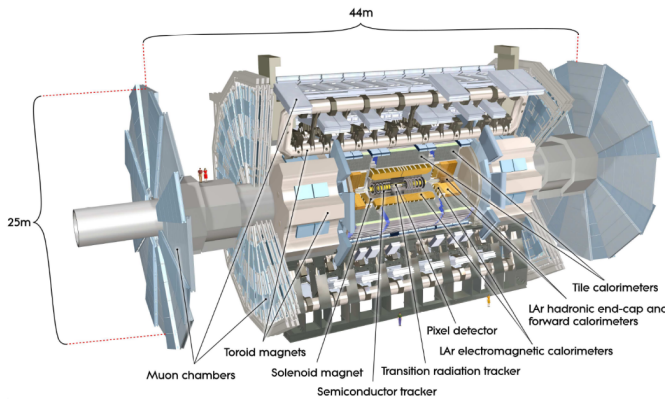


Figura 1: Representação 3D do detector ATLAS. Retirado de [15].

cada 25 nanossegundos, o volume de dados pode alcançar aproximadamente 70 TB/s, o que impede o armazenamento de todos os eventos [4]. Para armazenar apenas partículas de interesse, um sistema de filtragem *online* (*Trigger*) é empregado [10].

Elétrons em estados finais, ou seja, o conjunto de partículas detectadas após a colisão e quaisquer decaimentos subsequentes de partículas instáveis, estão relacionados com diferentes decaimentos físicos relevantes. Entretanto, a sua detecção torna-se complexa, pois sua faixa de energia é dominada por um ruído de fundo produzido por jatos hadrônicos (prótons), os quais geram um volume substancial de dados não pertinentes [16]. Neste contexto, o sistema de *trigger* do ATLAS é projetado de forma hierárquica, realizando filtragem para selecionar os eventos de interesse físico [17]. A supressão da maior parcela do ruído de fundo ocorre nos estágios iniciais de seleção, que empregam critérios mais simples, permitindo assim a aplicação de algoritmos mais sofisticados e computacionalmente intensos somente aos eventos candidatos a conter a física de interesse [16].

O programa de atualizações do LHC contempla um aumento na energia das colisões. Tais modificações impõem condições operacionais mais severas aos detectores, notadamente pelo aumento da taxa de eventos. Consequentemente, isso exige sistemas de filtragem de dados que sejam capazes de operar em tais condições [18]. No período de coleta de dados previsto para ocorrer entre 2030 - 2034 (*Run-4*) a energia de colisão aumentará em uma ordem de magnitude, e o número médio de interações por colisão aumentará para 200, produzindo um alto empilhamento de sinais. O ambiente operacional do *High Luminosity LHC* (HL-LHC) exige uma atualização do *Trigger and Data Acquisition* (TDAQ) – Sistema de seleção e aquisição de dados – do experimento ATLAS para manter eficiência de detecção semelhante, mesmo com taxas de eventos mais altas. Um novo componente de hardware, o *Global Trigger*, será adicionado ao novo TDAQ. Ele implementará a seleção de partículas em hardware digital, usando a tecnologia *Field Programmable Gate Array* (FPGA), a 1 MHz de taxa de eventos de saída. Os algoritmos de computação e seleção serão

incorporados ao *Global Event Processor* (GEP), o núcleo do *Global Trigger*.

Nesse cenário, este trabalho propõe uma solução de hardware para FPGA para implementar um método de identificação online de elétrons baseado em aprendizagem de máquina, chamado *NeuralRinger* [16]. O circuito proposto para a implementação do *NeuralRinger* foi construído como um componente, circuito, que será embarcado no *framework* do GEP.

O artigo está organizado em 5 seções. Na seção 2 é apresentado o sistema de *Trigger* do ATLAS que estará em operação na *Run-4* e o algoritmo do *NeuralRinger*, que foi implementado para FPGA neste trabalho. Na seção 3 é apresentada a arquitetura de hardware proposta para a implementação do *NeuralRinger* em hardware, considerando o ambiente de operação do *Global Trigger*. Os resultados obtidos com a simulação e a síntese lógica do circuito são apresentados na seção 4, e as conclusões são apresentadas na seção 5.

II. SISTEMA DE TRIGGER DO ATLAS PARA A RUN4

O sistema de trigger do ATLAS será atualizado para as novas condições de operação da *Run-4*. Um nível zero (L0) de identificação de eventos tem sido desenvolvido em hardware, e realizará a leitura dos sinais dos calorímetros para classificação dos eventos. Na Figura 2, é apresentada a arquitetura proposta para o TDAQ na *Run-4*. A janela de tomada de decisão do sistema de *trigger* deverá operar em até 10 μ s. O resultado da identificação dos eventos e os dados do detector são transmitidos através do Sistema de Aquisição de Dados (DAQ) a 1 MHz. O sistema Filtro de Eventos (EF - *Event Filter*) é composto por um conjunto de processadores que devem reduzir a taxa de eventos para 10 kHz. Os eventos selecionados pela decisão do EF são finalmente armazenados em mídia permanente.

A. *NeuralRinger*

Para a identificação de partículas eletromagnéticas, como elétrons e fótons, no sistema de *Trigger* do ATLAS, emprega-se um algoritmo de classificação conhecido como *NeuralRinger* [20]. A estratégia fundamental deste algoritmo reside na extração de características distintivas a partir dos sinais depositados no calorímetro, organizando-os segundo uma topologia de anéis concêntricos. Estes anéis são definidos com base no perfil espacial da deposição de energia da partícula incidente. As informações quantitativas de cada anel servem, então, como entrada para uma rede neural artificial. Esta rede é encarregada de realizar um teste de hipóteses, cujo resultado permite inferir a natureza da partícula detectada [21]. A arquitetura de rede neural utilizada para esta tarefa de classificação supervisionada é um Perceptron de Múltiplas Camadas (MLP) [22]. Na Figura 3 é ilustrado o diagrama que representa o funcionamento do *NeuralRinger* em duas etapas: *FEX* e *HYPO*.

A construção dos referidos anéis, um processo denominado “anelamento”, é iniciada pela identificação de Regiões de Interesse (*RoIs*) que circundam a célula de maior deposição

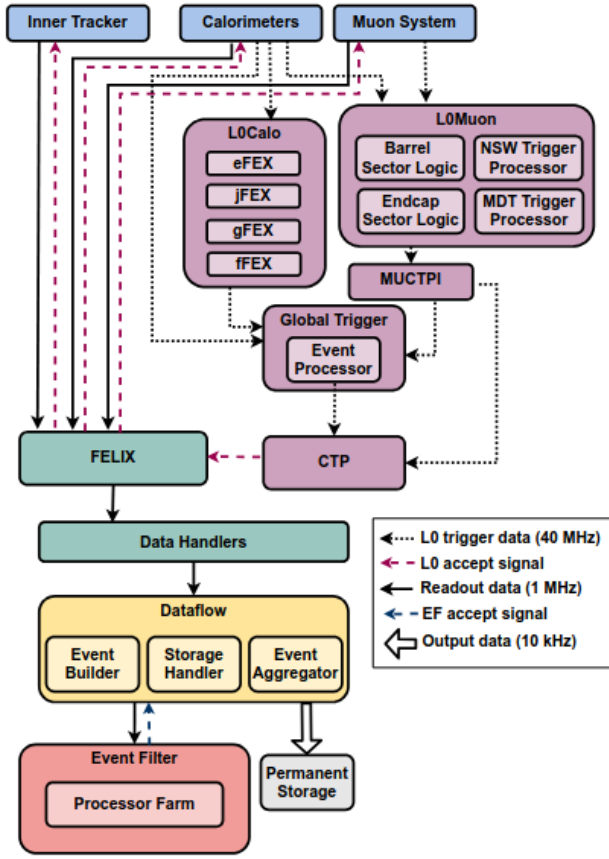


Figura 2: Diagrama de blocos do novo sistema de trigger do ATLAS para a Run-4 [19].

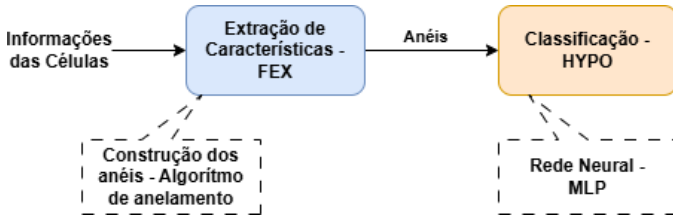


Figura 3: Diagrama de Blocos do NeuralRinger.

energética em cada uma das camadas do calorímetro. Esta célula central de maior energia estabelece o primeiro anel da estrutura. Sequencialmente, as células vizinhas a este primeiro anel são agrupadas para formar o segundo anel, e este padrão de agregação prossegue até que toda a extensão da *RoI* seja coberta por anéis concêntricos. Para cada anel assim formado, a energia total é calculada pela soma das energias detectadas nas células que o compõem, gerando um sinal de energia específico para aquele anel. Os diversos sinais de energia, provenientes dos anéis de todas as camadas do calorímetro, são subsequentemente agregados em um vetor de características único, que totaliza 50 anéis (exemplificado na Figura 5). Conforme ilustrado na figura, os calorímetros eletromagnético e hadrônico são segmentados em três diferentes camadas. *EM1* é a primeira camada do eletromagnético, *EM2* a segunda e

EM3 a terceira. *HAD1* é a primeira camada do hadrônico, *HAD2* a segunda e *HAD3* a terceira. Por fim, *PS* é a camada do Calorímetro Pré-Amostrador (PS) [16]. Um aspecto importante a ser destacado é a variação na granularidade intrínseca às diferentes camadas do calorímetro, o que resulta na geração de um número distinto de anéis para cada camada. Esta particularidade é ilustrada na Tabela I.

A Figura 4 apresenta o perfil médio dos anéis a partir de dados de simulação do ATLAS. O gráfico ilustra as diferenças esperadas no perfil do anel entre os elétrons e as partículas de jato hadrônicos.

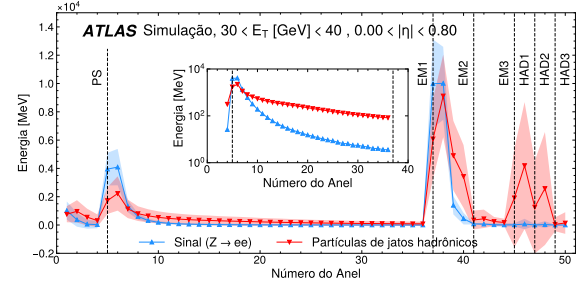


Figura 4: Gráfico do perfil médio dos anéis.

Tabela I: Número de anéis por camada do calorímetro. Fonte: [16].

Camada:	PS	EM1	EM2	EM3	HAD1	HAD2	HAD3
Anéis:	4	32	4	4	2	2	2

III. ARQUITETURA DE HARDWARE PROPOSTA

A implementação do algoritmo do *NeuralRinger* em hardware desenvolvida neste projeto foi projetada para ser embarcada no GEP, que é o módulo do sistema do *Global Trigger* no qual os algoritmos são incorporados [23]. O sistema de hardware do GEP utiliza a estrutura de FIFO (*First In, First Out*) para a transmissão de dados entre o framework e os circuitos dos algoritmos. FIFOs são utilizadas para armazenar as informações das células de cada evento e também o processamento gerado por cada algoritmo.

Na Figura 6 é apresentado o diagrama de blocos da implementação do *NeuralRinger* integrado ao GEP. O algoritmo do *NeuralRinger* utiliza informações das células do Calorímetro de Argônio Líquido (*Liquid Argon - LAr*) e do Calorímetro Hadrônico de Telhas (*Tile Calorimeter - TileCal*), desta forma, o circuito possui como entradas dados de três diferentes FIFOs, duas referentes às células do LAr (*LAr 1* e *LAr 2*) e uma referente às células do *TileCal*. O resultado do algoritmo, a decisão de *trigger*, é escrito em uma FIFO para ser utilizado pelos demais blocos da cadeia do sistema de *Trigger*.

O *NeuralRinger* é dividido em duas etapas e, conforme pode ser observado na Figura 3, cada etapa foi implementada como um bloco diferente: a etapa de extração de características

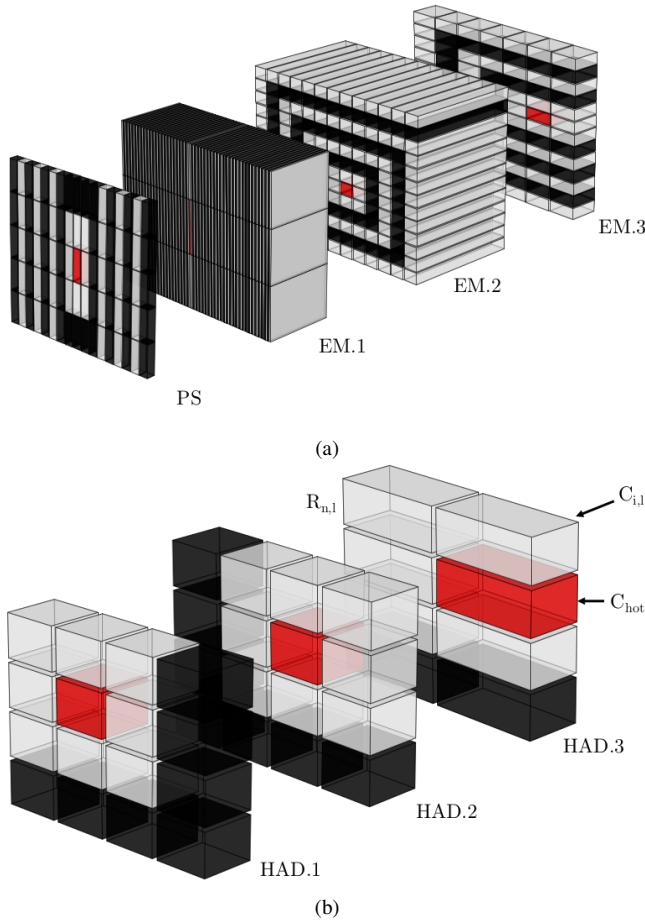


Figura 5: Ilustração do processo de anelamento para os calorímetros eletromagnético (a) e hadrônico (b). Fonte: [16].

(*Feature Extraction* - FEX) e de Classificação (*Hypothesis* - HYPO). Na primeira etapa, o circuito representado pelo bloco FEX lê os dados das células e constrói os 50 anéis, referentes as camadas do LAr e do TileCal. O circuito HYPO recebe os anéis e realiza a classificação, gerando a decisão de *trigger*. Essa decisão é a saída do circuito, que é gravada no FIFO de saída do GEP.

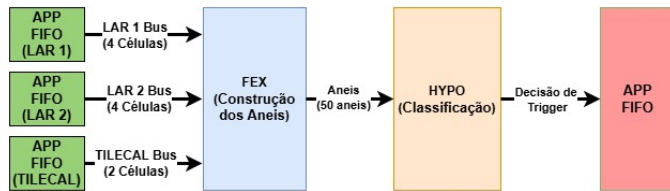


Figura 6: Diagrama de Blocos do NeuralRinger e a integração com o GEP.

A. Feature Extraction

O circuito *FEX* foi implementado em estágios de pipeline. O algoritmo de construção dos anéis é composto por diferentes cálculos matemáticos [16] e, cada operação matemática do

algoritmo foi implementada como um estágio do pipeline. A Figura 7 mostra o diagrama de blocos da implementação do circuito do *FEX*.

O *FEX* é o circuito responsável por construir os anéis a partir dos dados das células. Dessa forma, esse circuito possui interface com o GEP para acessar as FIFOs que armazenam as informações dos eventos. Neste circuito, para controlar a transmissão dos dados, foi implementada uma máquina de estado que realiza leituras sucessivas do FIFO até que todos os dados do evento tenham sido lidos. Em cada operação de leitura, um conjunto de células (4 células do LAr 1, 4 do LAr 2 e 2 do Tile) é lido e iniciado seu respectivo processamento no pipeline.

O primeiro estágio do pipeline é composto por dois circuitos: *Seed Finder* e *Separate Cell*. O primeiro é responsável por identificar e armazenar a semente utilizada pelo algoritmo como a célula mais quente, a partir da qual, os anéis são construídos. O segundo bloco é responsável por separar e armazenar em diferentes registradores as células transmitidas pelos barramentos (LAr 1, LAr 2 e TileCal).

O segundo e terceiro estágio são responsáveis pela implementação da etapa do algoritmo do *NeuralRinger* que define a qual anel a célula pertence. O *Calculate Distance* se aproveita da geometria cilíndrica do calorímetro e calcula as distâncias, em coordenadas azimutais (η) e cilíndricas (ϕ), entre as células e a semente. O *Calculate Index* define, a partir do valor máximo entre as distâncias η e ϕ , o índice do anel ao qual cada célula pertence.

O estágio seguinte é responsável por decodificar os valores de energia de cada célula e extrair as informações sobre a qual camada cada célula pertence. O bloco *Decode Cells* gera, para cada célula, os valores de energia decodificados, o índice do anel e a camada a qual cada célula pertence. O *framework* codifica os valores de energia de cada célula em 11 bits, desta forma, este bloco realiza a decodificação para 32 bits para que as operações de acúmulo sejam realizadas nos estágios seguintes.

Os últimos estágios do pipeline são responsáveis por acumular a energia de cada célula em seu respectivo anel. Em cada posição da FIFO são armazenadas 4 células, para o LAr 1 e LAr 2, e duas células para o TileCal. Desta forma, por leitura, até 4 células são processadas pelo pipeline em cada estágio. Para garantir o correto funcionamento da lógica de acúmulo das energias em seus respectivos anéis, a lógica para acumular a energia dessas células foi implementada em 4 estágios diferentes. Não existe uma organização das células que são enviadas para o algoritmo, podendo, por exemplo, todas as células de uma palavra da FIFO pertencerem ao mesmo anel ou a anéis distintos. O primeiro estágio acumula a energia da primeira célula, o próximo estágio acumula a energia da segunda célula e assim por diante. Depois que todas as células de cada evento tiverem sido lidas e processadas pelo pipeline, os anéis serão formados e estarão disponíveis na saída do último estágio dos estágios de acumulação.

No final de cada evento, com os 50 anéis construídos, o circuito *FEX* gera um sinal de controle e transmite os anéis

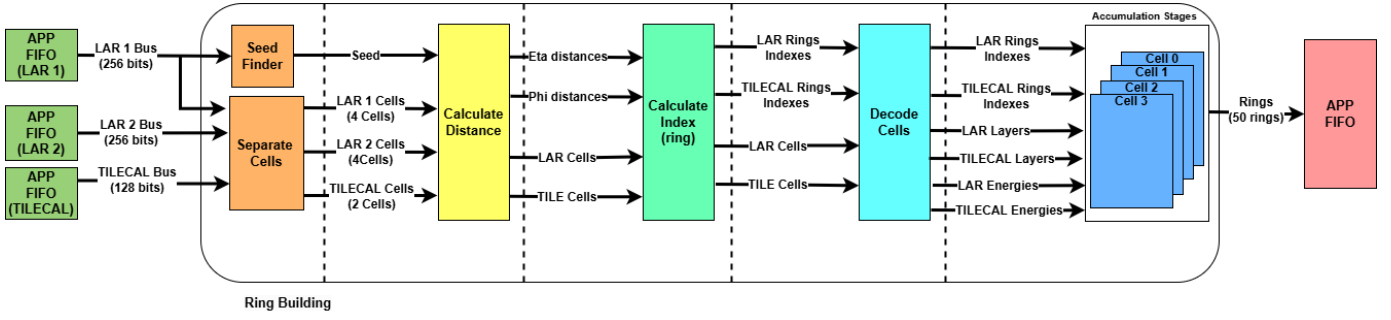


Figura 7: Diagrama de Blocos da Implementação do Circuito do FEX.

para o circuito do HYPO.

B. Hypothesis

A segunda parte do *NeuralRinger* é constituída de um conjunto de MLPs, uma para cada região do calorímetro. Embora difiram nos parâmetros, essas redes compartilham a mesma arquitetura: uma única camada oculta com cinco neurônios (utilizando a função de ativação ReLU) e um neurônio de saída. Este último realiza a classificação binária, indicando se um candidato a elétron é relevante ou não, por meio da função de ativação sigmóide. Essa classificação é feita a partir dos anéis construídos na FEX.

Essa topologia de rede foi selecionada após uma série de estudos comparativos envolvendo diferentes arquiteturas e configurações. O principal objetivo foi estabelecer um equilíbrio entre a eficiência na identificação de elétrons e a necessidade de execução extremamente rápida, uma vez que o algoritmo é operado no ambiente do *High-Level Trigger* (HLT) do ATLAS, que impõe severas restrições de latência. A configuração atual está em uso no sistema de *trigger* online desde 2017 [16].

Com o objetivo de reduzir o custo computacional do circuito implementado em Verilog, a MLP desenvolvida neste estudo foi projetada utilizando representação em ponto fixo no formato Q. As entradas foram codificadas em formato Q15 (16 bits), enquanto os pesos adotaram o formato Q7.15 (23 bits). A maior largura de palavra dos pesos foi necessária para acomodar operações intermediárias com maior precisão nos neurônios. A escolha desses formatos permitiu o alinhamento eficiente dos operadores aritméticos, simplificando a lógica de implementação dos neurônios em Verilog.

A Figura 8 apresenta um diagrama simplificado da arquitetura da rede neural proposta. Uma etapa fundamental do circuito é o bloco responsável pela atualização dos parâmetros da rede (pesos e bias). Esse bloco é implementado por meio de um registrador com 261 posições, destinado ao armazenamento dos parâmetros necessários para a classificação dos anéis provenientes de diferentes regiões do detector.

Outro desafio enfrentado no desenvolvimento em hardware foi a implementação da função sigmoide. Por envolver uma divisão e uma exponenciação, sua implementação direta em FPGA exigiria um elevado esforço tanto de desenvolvimento

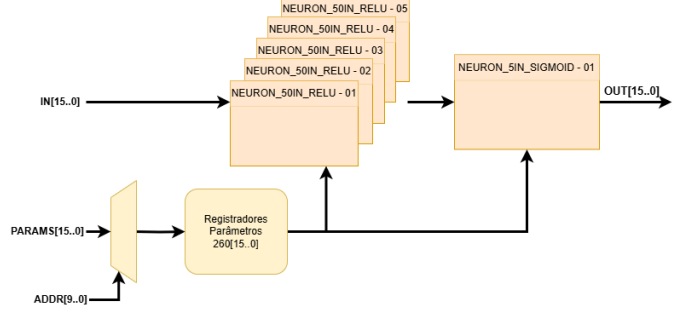


Figura 8: Diagrama de blocos do Hypo.

quanto de recursos computacionais. A utilização de *Look-Up Tables* (LUTs) foi desencorajada pela equipe de desenvolvimento do *framework* do GEP, uma vez que esse tipo de solução costuma consumir uma quantidade significativa de unidades lógicas no FPGA. Diante disso, tornou-se necessário buscar uma alternativa de implementação mais simples, com baixo custo computacional e menor ocupação de recursos do dispositivo.

Para contornar essa limitação, adotou-se uma abordagem baseada em uma aproximação linear por partes. Foram definidos intervalos no gráfico da função sigmoide, e, para cada intervalo, foi utilizada uma equação de reta que o aproxima. Dessa forma, a função é representada por um conjunto de expressões lineares simples, reduzindo significativamente a complexidade de implementação. O resultado dessa aproximação é apresentado na Figura 9.

IV. RESULTADOS

O circuito do *NeuralRinger* implementado foi simulado junto ao *framework* do GEP. O GEP possui uma estrutura de simulação em que é possível simular os circuitos dos algoritmos que serão embarcados considerando a arquitetura do *framework* em que os dados dos eventos são armazenados em FIFOs. Na Figura 10 é apresentado um diagrama de blocos que representa a estrutura de simulação implementada.

O ambiente de simulação do *framework* lê os dados dos eventos que serão utilizados na simulação a partir de arquivos texto. Desta forma, foi criado um código em Python que acessa uma base de dados de eventos do ATLAS e cria os arquivos de

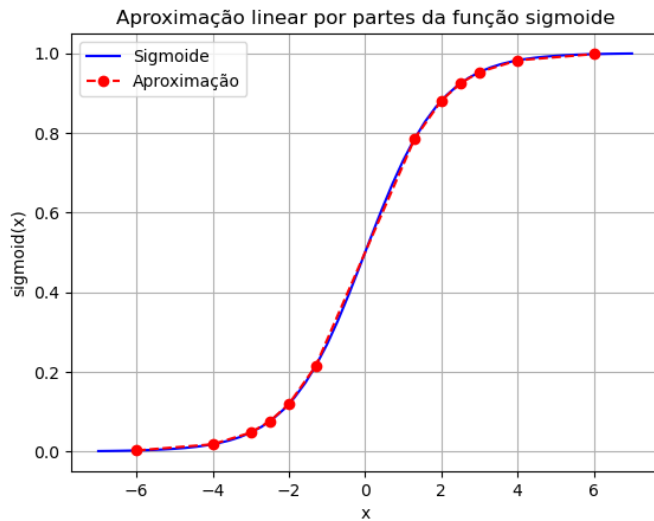


Figura 9: Aproximação linear por partes da função sigmoide.

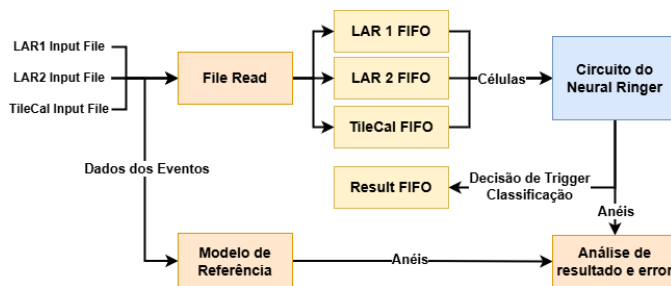


Figura 10: Diagrama de Blocos do Ambiente de Simulação.

entrada para o ambiente de simulação. Como o *NeuralRinger* recebe dados de LAR e do TileCal, foram necessários gerar três arquivos de entrada (dois para o LAR e um para o TileCal).

Para validação do circuito do *NeuralRinger*, foi construído, em software, um modelo de referência. Foram executadas múltiplas simulações com um conjunto de dados de cerca de 100 eventos e, nessas simulações, o circuito e o modelo de referência foram estimulados com o mesmo conjunto de dados de entrada.

Como o circuito do *NeuralRinger* opera somente com ponto fixo e os dados gerados pelos eventos do ATLAS são valores reais, para validar o funcionamento da etapa do FEX, foram realizadas duas análises de erros: uma considerando os anéis construídos pelo circuito e o modelo de referência com dados inteiros e outra comparando-os com dados reais gerados pelo modelo de referência. A análise do erro quadrático médio calculado com a saída do circuito e do modelo de referência simulados com valores inteiros mostraram que não houve erro, desta forma, verifica-se que o circuito apresentou o comportamento esperado quando comparado ao modelo de referência.

Para analisar a eficiência do circuito de construção dos anéis, o modelo de referência foi simulado com os valores reais dos eventos e o circuito com os valores representados em

inteiros. Com os resultados obtidos, foi construído o gráfico de barras de erro do erro médio normalizado e o desvio padrão, apresentado na Figura 11. No gráfico, é possível observar que o erro máximo é da ordem de 1,4%, indicando que a implementação do circuito é viável.

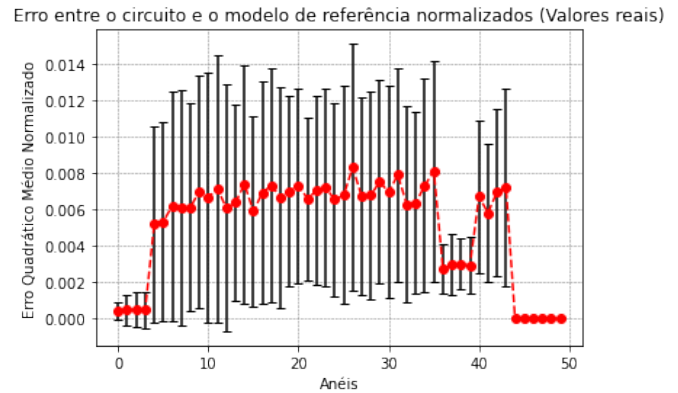


Figura 11: Erro quadrático médio (círculos vermelhos) e desvio padrão para cada anel, calculados para todos os eventos disponíveis, considerando a saída do circuito e da referência para valores reais.

Além da simulação funcional, também foi executada a síntese e gerados os resultados de ocupação de recursos lógicos da Placa FPGA. O projeto do GEP utiliza o FPGA Virtex UltraScale+ VCU118 [24], desta forma, a síntese do circuito de construção dos anéis foi sintetizada para esta placa. Na Tabela II são apresentados os resultados de utilização de *LookUp Tables (LUT)*, de estruturas de *LUT* configuradas como memórias *RAM*, as *Look-Up Table RAM (LUTRAM)*, de Flip-Flops (FF) e de blocos de memória *RAM* dedicadas, as *Block Random Access Memory (BRAM)*.

Tabela II: Utilização de recursos lógicos da síntese do circuito de construção dos anéis para a placa FPGA Virtex UltraScale+ VCU118

Recurso	Disponível	Estimativa de Utilização	Utilização %
LUT	1182240	16730	1.42
LUTRAM	591840	103	0.02
FF	2364480	22797	0.96
BRAM	2160	30.50	1.41

Para a verificação da rede neural na etapa da HYPO, foram utilizados os mesmos dados empregados durante o desenvolvimento e treinamento da MLP em software [16]. Para que esses dados pudessem ser utilizados na implementação em hardware, foi necessário convertê-los de ponto flutuante para ponto fixo no formato Q. Esse processo de quantização permitiu a adaptação dos dados ao domínio da lógica digital. Após o tratamento inicial, foi possível realizar a verificação funcional da rede em ambiente de simulação.

Para analisar a variação entre os resultados da rede implementada em hardware em relação à versão em software, decorrente de possíveis erros de quantização, foi calculado o erro entre as saídas de ambas as redes. A Figura 12 apresenta

o histograma desses erros. A partir dessa análise, observa-se que cerca de 58% dos erros ficaram abaixo de 0,01, sendo o maior erro encontrado igual a 0,042, registrado em apenas 3 amostras.

Além disso, considerando que os pesos da rede variam conforme a região do calorímetro, foi realizada uma comparação do erro quadrático médio (RMSE, *Root Mean Square Error*) para diferentes regiões testadas, conforme mostrado na Figura 13. Embora algumas regiões apresentem um desvio padrão ligeiramente maior, os valores de erro da rede em hardware em relação à implementação em software são pequenos, indicando boa preservação do desempenho da rede após a quantização.

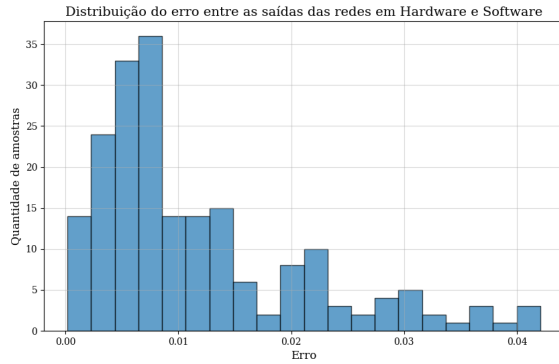


Figura 12: Histograma de distribuição do erro entre os valores de saída da rede em Hardware e Software

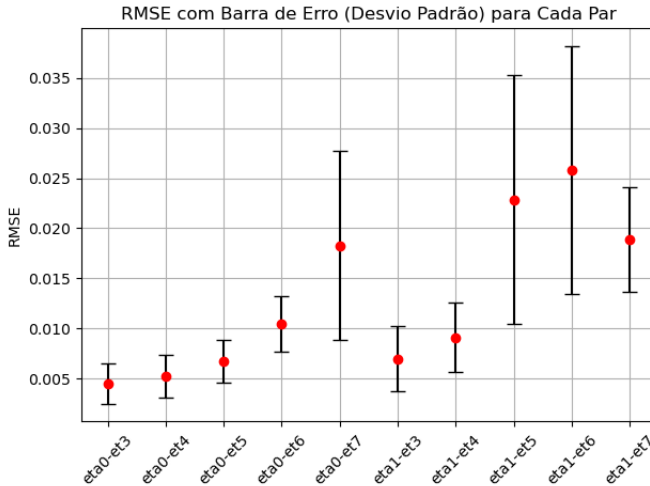


Figura 13: Erro Quadrático Médio entre as saídas da rede em Hardware e Software por região do calorímetro

Para a análise da utilização dos recursos lógicos, a rede neural implementada também foi sintetizada para a placa *FPGA Virtex UltraScale+ VCU118*. A Tabela III apresenta os dados obtidos com a síntese. Além dos recursos de *LUT* e *FF*, a síntese inferiu a utilização de estruturas de *Digital Signal Processors (DSPs)*, devido às operações aritméticas presentes no algoritmo da *MLP*.

Tabela III: Utilização de recursos lógicos da síntese do circuito da rede neural para a placa *FPGA Virtex UltraScale+ VCU118*

Recurso	Disponível	Estimativa de Utilização	Utilização %
LUT	1182240	1413	0.12
FF	2364480	6093	0.26
DSP	6840	265	3.87

V. CONCLUSÃO

Este trabalho apresentou a implementação em hardware (*FPGA*) do algoritmo do *NeuralRinger*, um método baseado em redes neurais, para seleção *online* de elétrons para o sistema *TDAQ* do experimento *ATLAS* do *LHC* proposto para *Run-4*. Os resultados obtidos ao longo deste trabalho demonstram um avanço significativo na implementação do algoritmo *NeuralRinger* em *FPGA*.

Foram apresentados os resultados dos circuitos de ambas as fases do algoritmo, a construção dos anéis (*FEX*) e a rede neural de classificação (*Hypo*). Ambos os circuitos apresentaram desempenho satisfatório, com erros da ordem de 1,4% para o circuito da *FEX* e 2% para o circuito da *Hypo* (decorrentes da quantização em ponto fixo), quando comparados aos resultados obtidos em ponto flutuante via software. Além disso, os resultados da síntese do circuito apontaram para uma baixa utilização dos recursos da placa *FPGA*, com 1,42% de utilização de *LookUp Tables* e 0,96% de *Flip-Flops* para a *FEX*, e 0,12% de *LookUp Tables* e 0,26% de *Flip-Flops* para a *Hypo*. Esses resultados indicam que o *NeuralRinger* é um algoritmo viável para ser embarcado no novo sistema de *Trigger* do *ATLAS*, no *Global Trigger*. Os resultados mostraram um desempenho adequado mesmo com as limitações inerentes ao ambiente de hardware.

Para as próximas etapas, é necessário prosseguir com a avaliação detalhada do desempenho do algoritmo embarcado, testando a rede neural com os anéis já construídos em hardware pela *FEX*. Além disso, será necessário conduzir um estudo aprofundado sobre técnicas de normalização aplicáveis aos anéis. O objetivo é identificar a estratégia de normalização mais adequada para ser implementada em *FPGA*, considerando restrições de recursos e eficiência.

Adicionalmente, já foi iniciado o processo de integração do *NeuralRinger* ao repositório do *framework* do *GEP*, com o objetivo de incluí-lo na lista de algoritmos utilizados pelo sistema de *trigger* global durante a *Run-4* do *LHC*.

AGRADECIMENTOS

Os autores agradecem ao CNPq, FAPESB, FAPERJ, Programa CAPES-COFECUB, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Código de Financiamento 001, LAA-HECAP e RENAFAE pelo apoio financeiro.

REFERÊNCIAS

- [1] W. N. Cottingham and D. A. Greenwood, *An Introduction to the Standard Model of Particle Physics*, 2nd ed. Cambridge University Press, Jul. 2023.
- [2] ATLAS Collaboration, "Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the LHC," *Physics Letters B*, vol. 716, no. 1, pp. 1–29, 2012. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S037026931200857X>

- [3] CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC," *Physics Letters B*, vol. 716, no. 1, pp. 30–61, 2012.
- [4] João Victor da Fonseca Pinto and ATLAS Collaboration, "An Ensemble of Neural Networks for Online Filtering Implemented in the ATLAS Trigger System," *Journal of Physics: Conference Series*, vol. 1162, p. 012039, March 2019, 18th International Conference on Calorimetry in Particle Physics (CALOR2018). [Online]. Available: <https://iopscience.iop.org/article/10.1088/1742-6596/1162/1/012039>
- [5] ATLAS Collaboration, "The ATLAS Experiment at the Large Hadron Collider," CERN, Tech. Rep. ATLAS-31-07-2019, July 2019, Overview document from ATLAS Rome Group. [Online]. Available: [https://www.roma1.infn.it/exp/atlas/](https://wwwroma1.infn.it/exp/atlas/)
- [6] "LHC Machine," *JINST*, vol. 3, p. S08001, 2008.
- [7] K. Aamodt and ALICE Collaboration, "The ALICE Experiment at the CERN LHC," *Journal of Instrumentation*, vol. 3, p. S08002, 2008.
- [8] A. A. Jr. and LHCb Collaboration, "The LHCb Detector at the LHC," *Journal of Instrumentation*, vol. 3, p. S08005, 2008.
- [9] S. Chatrchyan and CMS Collaboration, "The CMS Experiment at the CERN LHC," *Journal of Instrumentation*, vol. 3, p. S08004, 2008.
- [10] G. Aad and ATLAS Collaboration, "The ATLAS Experiment at the CERN Large Hadron Collider," *Journal of Instrumentation*, vol. 3, p. S08003, 2008.
- [11] A. Collaboration, "The ATLAS experiment at the CERN Large Hadron Collider: a description of the detector configuration for Run 3," *J. Instrum.*, vol. 19, no. 05, p. P05063, may 2024.
- [12] E. Ros, "ATLAS inner detector," *Nuclear Physics B - Proceedings Supplements*, vol. 120, pp. 235–238, 2003, proceedings of the 8th International Conference on B-Physics at Hadron Machines.
- [13] A. Airapetian *et al.*, "ATLAS calorimeter performance technical design report," ATLAS Collaboration, Tech. Rep., 12 1996.
- [14] S. Palestini, "The muon spectrometer of the ATLAS experiment," *Nuclear Physics B - Proceedings Supplements*, vol. 125, pp. 337–345, 2003.
- [15] T. Ciodaro, J. M. de Seixas, and A. Cerqueira, "Use of hadronic calorimetry information in the ATLAS level-1 muon trigger," *IEEE Transactions on Nuclear Science*, vol. 61, no. 2, pp. 1047–1055, 2014.
- [16] J. V. D. F. Pinto, "Filtragem Online Segmentada Baseada em Redes Neurais operando na informação de um calorímetro de altas energias de fina granularidade," Ph.D. dissertation, Universidade Federal do Rio de Janeiro - UFRJ, Apr. 2022, [Online]. Available: <http://www.pee.ufrj.br/index.php/pt/producao-academica/teses-de-doutorado/2022/2016033453-221/file>.
- [17] G. Aad *et al.*, "The ATLAS trigger system for LHC Run 3 and trigger performance in 2022," *Journal of Instrumentation*, vol. 19, p. P06029, June 2024. [Online]. Available: <https://iopscience.iop.org/article/10.1088/1748-0221/19/06/P06029>
- [18] V. A. Ferraz, "Detecção online eficiente de eventos raros utilizado detectores finamente segmentados," Ph.D. dissertation, COPPE/Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil, 2018, ph.D. dissertation.
- [19] "Technical Design Report for the Phase-II Upgrade of the ATLAS TDAQ System," CERN, Geneva, Tech. Rep., 2017. [Online]. Available: <https://cds.cern.ch/record/2285584>
- [20] J. M. Seixas, L. P. Calôba, M. N. Souza, A. L. Braga, and A. P. Rodrigues, "Neural second-level trigger system based on calorimetry," 1995. [Online]. Available: <https://cds.cern.ch/record/278995>
- [21] W. Freund, "Identificação de elétrons baseada em um calorímetro de altas energias finamente segmentado," Ph.D. dissertation, COPPE/UFRJ, Rio de Janeiro, Brazil, 2018, ph.D. Thesis in Electrical Engineering.
- [22] D. C. Teles, F. Batista, E. E. P. de Souza, E. F. S. Filho, P. C. M. A. Farias, and J. M. de Seixas, "Sistema neural segmentado de detecção online de elétrons utilizando pré-processamento estatístico," in *Anais do XX Congresso Brasileiro de Automática*, Belo Horizonte, MG, Brasil, September 2014, pp. 2493–2499, 20 a 24 de Setembro de 2014. [Online]. Available: <https://www.researchgate.net/publication/324788037>
- [23] M. Begel and D. Sankey, "Hardware Specifications for the Global Trigger," CERN, Geneva, Tech. Rep., 2021. [Online]. Available: <https://cds.cern.ch/record/2798389>
- [24] C. Dudley, S. Majewski, W. C. Fisher, J. R. Eastlack, P. Plucinski, F. Monticelli, L. Yao, T. M. Hong, and S. Tang, "ATLAS Trigger and Data Acquisition System Phase-II Upgrade: Global Trigger Event Processor Framework Initial Design Review Document," CERN, Geneva, Tech. Rep., 2020. [Online]. Available: <https://cds.cern.ch/record/2713130>