DESENVOLVIMENTO DE UM ACELERADOR INT8 BASEADO EM SYSTOLIC ARRAY EM FPGA INTEGRADO A UM SYSTEM-ON-CHIP

Thiago Fernandes

Itajaí – SC
2025

Thiago Fernandes

Desenvolvimento de um Acelerador INT8 Baseado em Systolic Array em FPGA Integrado a um System-on-Chip

Estudo acadêmico desenvolvido de forma independente como parte do aprofundamento em arquitetura de computadores, sistemas embarcados e aceleração de hardware em FPGA, no âmbito do curso de Engenharia da Computação da Universidade do Vale do Itajaí – UNIVALI.

Itajaí – SC
2025

RESUMO

O aumento da complexidade computacional em aplicações modernas tem impulsionado o uso de aceleradores de hardware, especialmente em áreas como processamento digital de sinais e aprendizado de máquina. Nesse contexto, dispositivos FPGA destacam-se por permitir a implementação de arquiteturas paralelas customizadas.

Este estudo apresenta o desenvolvimento de um acelerador de operações aritméticas em ponto fixo de 8 bits (INT8), baseado em uma arquitetura de systolic array e implementado em linguagem VHDL. O acelerador foi integrado a um sistema em chip por meio do barramento Avalon Memory-Mapped, utilizando um wrapper dedicado para encapsular a lógica de computação e permitir sua comunicação com um processador softcore.

Palavras-chave: FPGA. Acelerador de hardware. Systolic Array. System-on-Chip. VHDL.

SUMÁRIO

1 INTRODUÇÃO

O avanço das aplicações computacionais nas últimas décadas tem intensificado a demanda por maior desempenho e eficiência energética, especialmente em áreas como processamento digital de sinais, visão computacional e aprendizado de máquina. Muitos desses domínios envolvem operações matemáticas repetitivas e altamente paralelizáveis, que não são exploradas de forma eficiente por arquiteturas tradicionais baseadas exclusivamente em processadores de propósito geral.

Nesse contexto, aceleradores de hardware surgem como uma alternativa capaz de suprir tais limitações, delegando a execução de tarefas específicas a unidades dedicadas. Dispositivos FPGA destacam-se nesse cenário por permitirem a implementação de arquiteturas paralelas customizadas, possibilitando a exploração do paralelismo estrutural e temporal de forma eficiente.

Arquiteturas baseadas em *systolic arrays* têm sido amplamente utilizadas em aceleradores modernos devido à sua capacidade de reuso de dados, escalabilidade e eficiência computacional. Essas arquiteturas são particularmente adequadas para operações de multiplicação e acumulação, comuns em algoritmos de aprendizado de máquina.

Diante desse cenário, este estudo tem como objetivo o desenvolvimento de um acelerador de operações aritméticas em ponto fixo de 8 bits (INT8), baseado em uma arquitetura de systolic array, bem como sua integração em um sistema em chip utilizando FPGA. O trabalho aborda desde a implementação da lógica de computação em VHDL até a integração do acelerador em um SoC por meio do barramento Avalon Memory-Mapped, explorando conceitos de modularização, encapsulamento e integração hardware/software.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 ACELERADORES DE HARDWARE EM FPGA

Aceleradores de hardware são unidades especializadas projetadas para executar funções específicas de forma mais eficiente do que processadores de propósito geral. Ao delegar tarefas computacionalmente intensivas a aceleradores dedicados, é possível obter ganhos significativos de desempenho e eficiência energética.

Em dispositivos FPGA, aceleradores podem ser implementados diretamente na lógica reconfigurável, permitindo a criação de arquiteturas paralelas altamente customizadas. Essa flexibilidade possibilita a exploração de pipelines, paralelismo espacial e reuso de dados, características essenciais para aplicações de alto desempenho.

2.2 ARQUITETURA SYSTOLIC ARRAY

A arquitetura *systolic array* é composta por um conjunto de unidades de processamento interconectadas, nas quais os dados fluem de forma sincronizada entre os elementos. Cada unidade executa operações simples, enquanto a coordenação global do fluxo de dados resulta em elevado paralelismo e eficiência computacional.

Essa arquitetura é amplamente utilizada em aceleradores de aprendizado de máquina, especialmente em operações de multiplicação e acumulação, como em multiplicações matriciais e convoluções. O modelo de fluxo rítmico de dados reduz acessos à memória externa e favorece o reuso de informações intermediárias.

2.3 ARITMÉTICA EM PONTO FIXO INT8

O uso de aritmética em ponto fixo de 8 bits (INT8) tem se tornado comum em aceleradores de aprendizado de máquina devido à redução de área, consumo de energia e latência, quando comparado à aritmética em ponto flutuante. Apesar da menor precisão, técnicas adequadas de acumulação e quantização permitem manter resultados satisfatórios em diversas aplicações.

3 METODOLOGIA

O desenvolvimento deste estudo foi conduzido de forma incremental, partindo da definição da arquitetura de computação até sua integração em um sistema em chip completo. Inicialmente, foi projetada a unidade elementar de processamento responsável pelas operações de multiplicação e acumulação em formato INT8.

Em seguida, múltiplas unidades foram interconectadas em uma arquitetura de systolic array unidimensional, permitindo a exploração de paralelismo estrutural. Após a validação funcional do acelerador em nível RTL, foi desenvolvido um wrapper compatível com o barramento Avalon Memory-Mapped, com o objetivo de encapsular a lógica de computação e permitir sua comunicação com o software.

Por fim, o acelerador foi integrado em um sistema em chip utilizando o Intel Platform Designer, juntamente com um processador softcore, memória on-chip, fonte de clock e controlador de reset. Essa abordagem permitiu avaliar o funcionamento do acelerador em um ambiente próximo a um sistema embarcado real.

4 DESENVOLVIMENTO DO ACELERADOR

O acelerador desenvolvido neste estudo foi implementado integralmente em linguagem VHDL e estruturado de forma modular, visando facilitar sua compreensão, reutilização e expansão futura. A arquitetura adotada baseia-se em unidades elementares de processamento interconectadas em um systolic array unidimensional.

Cada unidade elementar de processamento realiza operações de multiplicação entre um dado de entrada e um peso, ambos representados em formato INT8. O resultado da multiplicação é acumulado em um registrador de 32 bits, garantindo segurança contra overflow durante operações sucessivas. A unidade é síncrona e utiliza pipeline para permitir operação em frequências mais elevadas.

O systolic array é formado pela interconexão sequencial dessas unidades, permitindo que os dados de entrada sejam propagados entre os elementos enquanto os pesos permanecem estacionários. Essa abordagem reduz a necessidade de acessos à memória e favorece o paralelismo computacional. O resultado final da acumulação é disponibilizado na saída do array para leitura pelo processador.

5 ARQUITETURA DO SYSTEM-ON-CHIP

A arquitetura do sistema em chip desenvolvido neste estudo segue um modelo clássico de SoC baseado em barramento, no qual um processador central coordena a execução das aplicações e o controle dos periféricos especializados. O acelerador INT8 foi integrado como um periférico customizado, acessível por meio do barramento Avalon Memory-Mapped.

A Figura 1 apresenta uma visão geral da arquitetura do sistema.

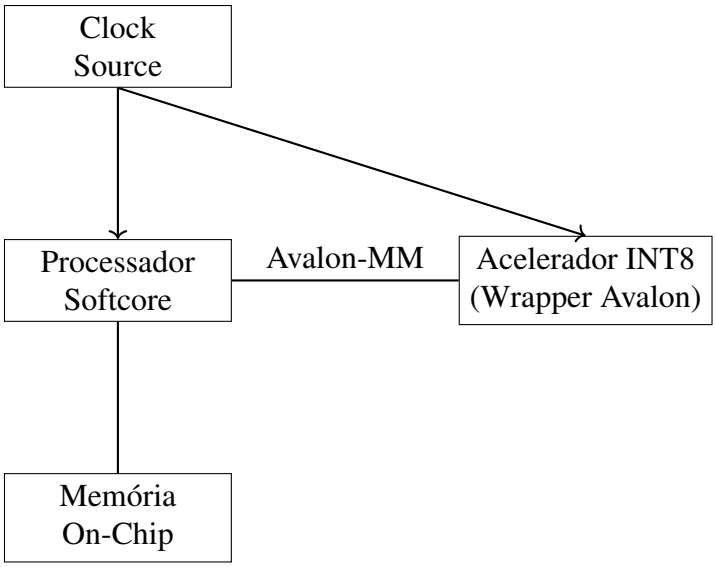

Figura 1 – Arquitetura do System-on-Chip com acelerador INT8 integrado

5.1 COMPONENTES DO SOC

O SoC é composto pelos seguintes blocos principais:

- **Fonte de Clock**: responsável pela geração do sinal de clock global do sistema.
- **Processador Softcore**: executa o software de controle e realiza acessos ao barramento Avalon-MM.
- **Memória On-Chip**: armazena o código do programa e dados de execução.
- **Controlador de Reset**: garante a inicialização síncrona e consistente dos módulos.
- **Acelerador INT8**: periférico customizado responsável pelas operações de multiplicação e acumulação.

5.2 MODELO DE COMUNICAÇÃO

A comunicação entre o processador e o acelerador ocorre exclusivamente por meio de acessos ao barramento Avalon Memory-Mapped. O acelerador é mapeado em um espaço de endereçamento específico e controlado por registradores, permitindo uma integração simples e eficiente.

Essa abordagem favorece a modularidade do sistema, pois o acelerador pode ser substituído ou expandido sem a necessidade de alterações na arquitetura global do SoC.

5.3 FLUXO DE DADOS DO SISTEMA

O fluxo de dados do sistema inicia-se no software executado pelo processador, que escreve os valores de entrada e os pesos nos registradores do acelerador por meio de acessos de escrita no barramento Avalon-MM. Após a configuração dos dados, o software ativa o acelerador por meio de um registrador de controle.

Internamente, o wrapper converte os acessos do barramento em sinais de controle e dados para o systolic array. Os dados de entrada são propagados sequencialmente entre as unidades de processamento, enquanto os pesos permanecem estacionários em cada elemento do array. O resultado acumulado é então disponibilizado em um registrador de saída, que pode ser lido pelo processador.

Esse modelo de operação permite desacoplar completamente o software da implementação interna do acelerador, reforçando o encapsulamento e a reutilização do hardware.

6 CONCLUSÃO

O estudo demonstrou a viabilidade da implementação de um acelerador INT8 baseado em systolic array em FPGA, bem como sua integração em um sistema em chip. A abordagem adotada possibilita a expansão futura da arquitetura e sua adaptação para diferentes aplicações.

REFERÊNCIAS

[1] INTEL CORPORATION. *Avalon Interface Specifications*. Intel FPGA Documentation.

[2] INTEL CORPORATION. *Platform Designer User Guide*. Intel FPGA Documentation.

[3] KURODA, T. et al. *Systolic Array Architectures for Machine Learning*. IEEE.