Universidade de São Paulo
Instituto de Matemática e Estatística
Bachalerado em Ciência da Computação

João Henrique Luciano

Implementação de algoritmos em FPGAs usando síntese de alto nível

São Paulo
Dezembro de 2018

Implementação de algoritmos em FPGAs usando síntese de alto nível

Monografia final da disciplina
MAC0499 – Trabalho de Formatura Supervisionado.

Supervisor: Prof. Dr. Alfredo Goldman

São Paulo
Dezembro de 2018

Resumo

O presente trabalho visa o estudo e experimentação do uso de ferramentas de síntese de alto nível para o design de circuitos eletrônicos. O intuito é destacar tecnologias emergentes de hardware, como os FPGAs, assim como aproximar engenheiros de software do desenvolvimento de hardware. Para tanto, foi usado o LegUp, um arcabouço de código aberto para síntese de alto nível em FPGAs. Logo, o estudo do funcionamento interno e uso dessa ferramenta foi primordial para a elaboração do projeto. Com ele, dois algoritmos foram desenvolvidos na linguagem de programação C e, dessa forma, efetuar-se a síntese de alto nível do algoritmo para um circuito descrito em Verilog e implementado em FPGA. Como efeito secundário, foi criada uma referência em português para o estudo e uso da síntese de alto nível usando o sistema operacional Ubuntu, a placa SoC FPGA Cyclone V, da Altera, e o arcabouço de código aberto LegUp.

Palavras-chave: FPGA, síntese de alto nível, hardware, algoritmos, C, Verilog.

Abstract

This work aims the study and usage of high-level synthesis tools for electronic circuits design. Alongside that, there is emphasis in newer technologies, such as FPGAs, as well as bringing together software engineers and hardware development. To do that, LegUp, a open-source high-level synthesis framework, is used along with a system-on-a-chip FPGA. Furthermore, the internal structure and usage of the framework was essential for the project development. Using LegUp, two algorithms were implemented using C programming language to feed its high-level synthesis process and create a Verilog HDL, which was downloaded into a FPGA. As side effect, there is the creation of a brazilian reference for this area of study, which is quite scarse.

Keywords: FPGA, high-level synthesis, hardware, algorithms, C, Verilog.

Sumário

Capítulo 1

Introdução

Uma monografia deve ter um capítulo inicial que é a Introdução e um capítulo final que é a Conclusão. Entre esses dois capítulos poderá ter uma sequência de capítulos que descrevem o trabalho em detalhes. Após o capítulo de conclusão, poderá ter apêndices e ao final deverá ter as referências bibliográficas.

Para a escrita de textos em Ciência da Computação, o livro de Justin Zobel, *Writing for Computer Science* (Zobel, 2004) é uma leitura obrigatória. O livro *Metodologia de Pesquisa para Ciência da Computação* de Wazlawick (2009) também merece uma boa lida.

O uso desnecessário de termos em lingua estrangeira deve ser evitado. No entanto, quando isso for necessário, os termos devem aparecer *em itálico*.

```
Modos de citação:
indesejável: [AF83] introduziu o algoritmo ótimo.
indesejável: (Andrew e Foster, 1983) introduziram o algoritmo ótimo.
certo : Andrew e Foster introduziram o algoritmo ótimo [AF83].
certo : Andrew e Foster introduziram o algoritmo ótimo (Andrew e Foster, 1983).
certo : Andrew e Foster (1983) introduziram o algoritmo ótimo.
```

Uma prática recomendável na escrita de textos é descrever as legendas das figuras e tabelas em forma auto-contida: as legendas devem ser razoavelmente completas, de modo que o leitor possa entender a figura sem ler o texto onde a figura ou tabela é citada.

Apresentar os resultados de forma simples, clara e completa é uma tarefa que requer inspiração. Nesse sentido, o livro de Tufte (2001), *The Visual Display of Quantitative Information*, serve de ajuda na criação de figuras que permitam entender e interpretar dados/resultados de forma eficiente.

Capítulo 2

FPGA

2.1 Introdução

*FPGA*s (do inglês *Field Programmable Gate Array*) são dispositivos de silício que podem ser programados após sua fabricação, permitindo que quase qualquer *design* de circuito digital possa ser implementado nele. Essa maleabilidade torna-os atrativos para tarefas que envolvam a produção e alteração de *designs* de circuitos lógicos, pois o custo e tempo de execução dessas atividades são muito reduzidos em comparação ao tradicional ciclo de desenvolvimento com o uso de dispositivos *ASIC*s (do inglês *Application-Specific Integrated Circuit*).

2.2 Reprogramabilidade

A flexibilidade do *FPGA* se deve à sua capacidade de ser reprogramada após sua fabricação. Para tanto, interruptores programáveis são posicionados em suas rotas de interconexão. Como esses interruptores (também conhecidos como *switch*s) são componentes eletrônicos que redirecionam a corrente elétrica, é necessário que o interruptor seja configurado para que cada corrente de entrada seja desviada para a rota certa. Isso é feito a partir de memórias, que são programadas juntas do FPGA ao subir um programa no *chip* ou na placa que o contém.

Além das rotas de interconexão, as portas lógicas do FPGA também devem ser arranjadas de tal forma que o processamento dos dados seja feito de acordo com o especificado pelo usuário. Ao invés de criar um possível esquema de tradução da linguagem de especificação de hardware para funções booleanas, são empregadas as chamadas *LUTs* (*lookup tables*).

Para definir qual o circuito implementado no *chip*, certas tecnologias de programação de circuitos são usadas, memórias para guardar os estados dos *switchs* e de outros componentes. Idealmente, essas memórias são não-voláteis, infinitamente reprogramáveis, baratas e que consumam pouca energia. Atualmente, a tecnologia mais utilizada é a *SRAM*, que não atinge todos os critérios ideais mas suas desvantagens podem ser facilmente contornadas.

A *SRAM* (*'Static Random Access Memory'*, ou 'memória estática de acesso aleatório') é um tipo de memória volátil que utiliza uma combinação de 6 transístores para guardar um *bit* de informação. O termo 'estático' refere-se à falta de necessidade de atualizações de memória (*memory refreshes*) em seu circuito. Note que isso não significa que, na ausência de corrente elétrica, o estado se manterá: caso haja uma queda de energia, a memória perderá o dado nela contido.

Devido à sua característica estática, as *SRAM*s são mais rápidas e consomem menos

energia que as *DRAM*s. Entretanto, como as *SRAM*s necessitam de 6 transistores, ao contrário das *DRAM*s que utilizam apenas 1 transistor e 1 capacitor, aquelas são muito mais caras e demandam mais espaço no circuito para serem implementadas, além de ambas serem voláteis.

2.3 Componentes

A composição de um *FPGA* pode ser resumido em três componentes:

- Blocos lógicos
- Rotas de interconexão
- Blocos de entrada e saída

Os blocos lógicos são configuráveis e implementam funções lógicas e armazenamento de dados (i.e. memória). Os blocos de E/S recebem e enviam dados para fora do *chip*. Por fim, as rotas de interconexão conectam os blocos lógicos entre si e entre os blocos de E/S. Uma forma de visualizar esses componentes é através de uma matriz, onde os blocos lógicos estão dispostos bidimensionalmente, numa grade, e conectados pelas rotas de interconexão. Nas bordas dessa matriz se encontram os blocos de E/S, integrados à matriz pelas rotas de interconexão, servindo para a comunicação do *FPGA* com dispositivos exteriores a ele.

2.3.1 Blocos lógicos

Os blocos lógicos configuráveis (BLCs) são as unidades que provêm capacidade lógica e de armazenamento para o FPGA. BLCs podem ser implementados de diversas maneiras, desde simples transistores até processadores inteiros, e essa implementação define sua granularidade. BLCs com granularidade muito pequena, como transistores, ocupam muito espaço e os torna ineficientes em termos de área. Por outro lado, os de granularidade muito grande, como processadores, podem representar um desperdício de recurso quando tratamos de funções mais simples. Entre esses máximos, temos um espectro de implementações de BLCs.

Os BLCs são compostos blocos lógicos básicos (BLBs), que podem ser usados em conjunto ou de forma isolada, ou seja, um BLC pode ser composto por um único BLB ou por um conjunto deles. As componentes usadas noss BLBs podem variar, mas a fabricante da placa usada no presente trabalho, a Altera, utiliza *lookup tables* e *flip-flops* para armazenamento. As LUTs são usadas como tabela de valor para representar qualquer função booleana com determinado número de *bits* de entrada e de saída. Assim, uma LUT que recebe k *bits* é chamada de *k-LUT* e representa qualquer função booleana f tal que

$$f : \{0,1\}^k \to \{0,1\}^k$$

A vantagem do uso de LUTs e *flip-flops* reside em não ter uma granularidade nem muito pequena, nem muito grande, permitindo o uso em conjunto para implementações mais complexas.

Apesar do uso de LUTs ou outros métodos para implementar BLBs, como NANDs, esses métodos são mais usados para criar a parte programável do FPGA. Uma parte dele pode vir já programada com blocos lógicos especializados, como processadores de sinais digitais (conhecidos como DSPs, *digital signals processor*), multiplicadores, somadores, ALUs

inteiras, todos criados de forma otimizada para suas tarefas. Estes são chamados de blocos rígidos, pois não podem ser reprogramados, apenas usados como estão no FPGA. Isso implica em um possível desperdício de espaço e recursos no caso desses blocos não serem utilizados pelo circuito, mas também traz a vantagem de se usar blocos dedicados a determinadas tarefas.

2.3.2 Rede de interconexão

Como já mencionado, a flexibilidade de um FPGA vem da capacidade de ter sua rede de interconexão reprogramada. Essa rede precisa ser flexível não só em termos de configuração de rotas, mas também de tipos de fios presentes no dispositivo para poder implementar uma grande variedade de circuitos. Apesar da maior parte das componentes de um circuito apresentar localidade (isto é, se localizarem perto umas das outras), há conexões que podem necessitar de fios mais longos. Cerca de 85% da área de um FPGA consiste da rede de interconexão entre os blocos lógicos. Visando otimizar a comunicação de acordo com a finalidade do circuito, essa rede pode ser construída usando arquiteturas diferentes. A arquitetura utilizada neste trabalho é a baseada em malha.

As redes baseadas em malha (do inglês '*mesh-based*', também conhecida como '*island-style*') são organizadas em formato matricial, com blocos lógicos cercados de fios de conexão - por isso o termo 'estilo de ilha', onde os BLs parecem estar "ilhados"em um "mar de fios". Nas extremidades se encontram os blocos de entrada e saída. Na rede de conexão estão localizadas concentrações de *switches* que estabelecem a rota dos sinais entre os blocos lógicos. Por último, existem as conexões entre os blocos lógicos e a rede de comunicação, que são chamadas de caixas de comunicação, também configuráveis. A figura 2.1 mostra essa arquitetura.

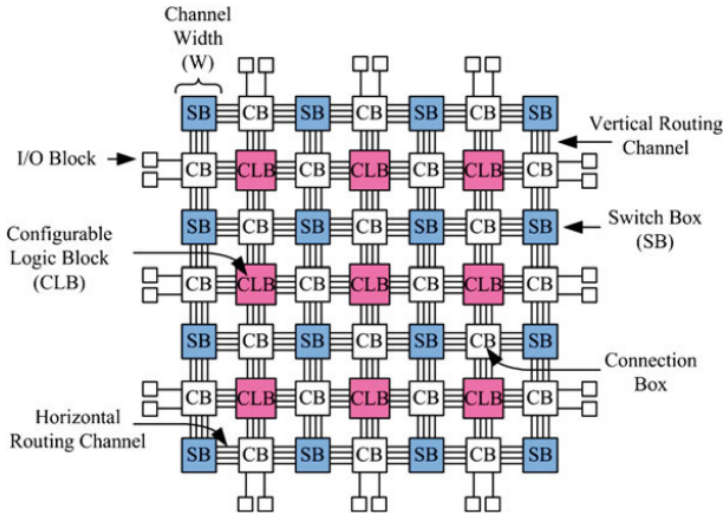

Figura 2.1: *Exemplo de um FPGA com rede de interconexão em malha.*

2.4 Desvantagens

A maior vantagem de *FPGAs* - sua flexibilidade - também é a causa de sua maior desvantagem. Essa característica baseia-se na reprogramação das rotas de interconexão, dos blocos lógicos e dos blocos de entrada e saída. Entretanto, a área usada por tais rotas

ocupa a maior parte do dispositivo, chengando a quase 90% da área útil do dispositivo, para permitir a sua reprogramação. Não obstante, os *switches* e os componentes necessários para implementar as LUTs geram resistência elétrica à propagação dos pulsos de *clock* do sistema, o que obriga os fabricantes a diminuir a frequência máxima de *clock* do *chip* e da placa. Por consequência disto, os FPGAs são mais lentos e consomem mais energia do que os *ASIC*s.

Um exemplo de figura está na figura 4.1.

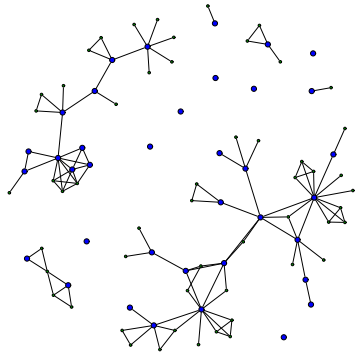

Figura 2.2: *Exemplo de uma figura.*

Capítulo 3

Síntese de alto nível

Síntese de alto nível ('High level synthesis', ou HLS) é o processo de transformação linguagens de programação de alto nível para sintetizar arquiteturas RTL ('*Register-transfer level*' ou 'nível de registrador e transferência'), isto é, sintetizar arquiteturas de circuitos digitais síncronos a partir de descrições comportamentais (ou algorítmicas) do *hardware*. As saídas geradas são, em sua maioria, arquivos em linguagens HDL ('*Hardware description language*' ou linguagem de descrição de *hardware*), que são usadas para configurar o *hardware*. Tais ferramentas realizam o mesmo fluxo básico na sintetização dos circuitos, desde a especificação comportamental desejada até a geração dos arquivos em HDLs.

3.1 Fluxo de síntese

O fluxo de síntese das ferramentas de HLS segue o mesmo padrão, que envolve:

- Modelagem ou especificação
- Compilação
- Alocação
- Escalonamento
- emparelhamento
- Geração

3.2 Modelagem

A etapa de modelagem consiste em descrever o comportamento desejado do *hardware* em questão, onde as entradas, saídas e forma de processamento dos dados são especificadas através do uso de linguagens programação de alto nível. As ferramentas de HLS, tais como o OpenCL(https://www.khronos.org/opencl/) e o LegUp(http://legup.eecg.utoronto.ca/), comumente usam linguagens com sintaxes baseadas em C.

A descrição é dita comportamental ou atemporal, onde o comportamento descrito recebe todos os dados de entrada simultaneamente, realiza seu processamento de forma instantânea, e retorna todos os dados de saída de uma vez. Esse tipo de descrição não é condizente com a forma como um sistema de *hardware* funciona, então há a necessidade de conversão do tipo atemporal para temporal.

É importante ressaltar que nem todos os algoritmos podem ser descritos diretamente em *hardware*. Um bom exemplo são algoritmos recursivos, que não são convertidos para formas iterativas de maneira automatizada.

3.3 Compilação

A modelagem atemporal deve ser compilada em outra temporal, onde os ciclos de *clock* do circuito são levados em consideração na execução das operações descritas. Para tanto, um modelo formal do comportamento do circuito é criado para visualizar melhor as dependências de dados e de controle de fluxo do algoritmo.

O modelo é representado por um grafo direcionado, chamado DFG ('*Data flow graph*' ou grafo de fluxo de dados) onde os arcos são valores constantes ou de variáveis e os vértices são operações que usam os valores. Essa forma de representação explicita o paralelismo intrínseco ao algoritmo descrito, facilitando as fases seguintes da síntese. Como os DFGs representam apenas fluxos de dados, há dificuldades em utilizá-los para representar laços limitados por variáveis ao invés de constantes (i.e `for (int i = 0; i < n; i++)`) ou trechos condicionais (`if-elses`). Para tanto, seria necessária transformações no grafo que, dependendo da complexidade da implementação, poderia gastar muito mais memória para armazenamento e muito mais processamento.

Pensando nisso, uma versão estendida do DFG foi criada, chamada CDFG ('*Control and Data Flow Graph*' ou grafo de fluxo de controle e dados), onde os arcos são controles de fluxo (como 'if-else's e 'goto's) e os nós são chamados de *blocos básicos*. Blocos básicos são blocos de código sequenciais sem pontos de saída (por exemplo, um 'return' ou 'goto' no meio do código) ou ramificações (causadas por fluxos condicionais, por exemplo) (podemos dizer que blocos básicos são os blocos de código entre um 'jump target' e um 'jump'). Os CDFGs são mais expressivos por conseguirem representar tanto o fluxo de dados quanto o de controle; entretando, faz-se necessária uma análise mais profunda para explorar o paralelismo dentro dos blocos básicos e expor o parelelismo entre os blocos.

3.4 Alocação

A compilação do modelo comportamental explicita as operações feitas no algoritmo e em qual ordem devem ser feitas. Após essa etapa, é preciso transformar essas representações abstratas no modelo lógico/físico do circuito.

Na fase de alocação, ocorre a identificação dos recursos de *hardware* necessários para implementar o circuito desejado. Dentre esses recursos, podemos citar as unidades funcionais, unidades de memória, de transferência etc. A alocação destes é feita usando a biblioteca RTL das ferramenta de HLS. A biblioteca contém os recursos disponíveis para cada modelo de *hardware*, bem como dados sobre esses recursos (e.g. área necessária, consumo de energia, latência), necessários para outras fases da síntese.

Vale lembrar que certas componentes a serem alocadas, principalmente as de comunicação (como os barramentos), podem ser deixadas para serem alocadas posteriormente a fim de otimização, como depois da fase de emparelhamento (para otimizar as comunicações entre as unidades funcionais) ou da fase de escalonamento (para não introduzir restrições de paralelismo entre as operações das unidades funcionais).

3.5 Escalonamento

Como apontado anteriormente, o processo de HLS transforma uma descrição atemporal para uma temporal, ou seja, que considera os ciclos de *clock* do sistema de *hardware*. A fase de escalonamento se encarrega de planejar o processamento dos dados de entrada a cada um desses ciclos, levando em consideração os dados especificados, as operações, as restrições desejadas do modelo (tais como área ou consumo de energia máximos) e os componentes alocados.

Durante essa fase, a representação do modelo em um CDFG é de extrema valia. Isso porque com o CDFG, evidencia-se o paralelismo entre blocos básicos (e, dependendo da profundidade da análise, o paralelismo dentro deles), e este é aproveitado pelo escalonador para otimizar o processamento de dados dentro das restrições estabelecidas. Aproveitam-se possíveis falta de dependência entre dados para realizar múltiplas operações por ciclo de *clock*, sob a restrição de haver unidades funcionais suficientes para tal (vê-se aí, por exemplo, a relação entre aumentar a área implementada de circuito implementada, número de recursos alocados e energia consumida, e diminuir o consumo de tempo e aumentar a taxa de processamento de dados). Dessa forma, uma operação pode ser escalonada pra ser executada ao longo de um ou mais ciclos de *clock*.

É também durante essa fase que pode ocorrer a comunicação entre a alocação e o emparelhamento para otimizar aspectos do *layout* do circuito digital, pois estas 3 fases estão intimamente ligadas por lidarem com o circuito de fato (diferente da modelagem e compilação, que lidam com o comportamento de forma ainda abstrata).

3.6 Emparelhamento

Para cada operação que nosso algoritmo descreve, precisamos não só alocar os recursos necessários para efetuar a operação como também ligar tanto as operações quanto as variáveis aos recursos alocados. A fase de emparelhamento (do inglês *binding*) é a responsável por essa tarefa, utilizando-se dos resultados das outras fases para fazer tais ligações. Nela, podem ocorrer mais otimizações (usufruindo da comunicação com as fases de escalonamento e alocação, como já citado) para diminuir a área utilizada, por exemplo: se uma mesma operação é feita em ciclos diferentes, pode-se reutilizar a unidade funcional daquela operação. Da mesma forma, unidades de memória podem guardar valores de variáveis que possuem um tempo de vida diferente.

3.7 Geração

Finalmente, após o algoritmo de síntese ter realizado todas as suas operações, é gerado um arquivo com uma arquitetura RTL representando o comportamento descrito pelo modelo. O arquivo de saída pode ser de diversos formatos, tais como SystemC, Verilog e VHDL. Vale ressaltar que cada *framework* geralmente trabalha com um número limitado de modelos de placa FPGA, uma vez que estão cada vez mais frequentes o uso de FPGAs em placas integradas, tornando-se um SoC FPGA (do inglês "*System-on-a-Chip FPGA*"). Mais informações estão disponíveis no capítulo 2, sobre FPGAs.

3.8 Alocação, Escalonamento e emparelhamento: considerações especiais

As fases de alocação, escalonamento e emparelhamento estão intimamente ligadas, como já observado ao longo da seção anterior. A compilação do programa e a geração do RTL transformam, respectivamente, linguagens de programação em uma representação intermediária e vice-versa. Já essas três fases manipulam a representação intermediária com o objetivo de dizer, de forma concreta, de quais recursos do *chip* e quando o algoritmo precisará deles. Essas etapas podem ocorrer de forma concorrente ou sequencial, dependendo da arquitetura da ferramenta, e essas formas de execução podem alterar a construção do circuito. Por exemplo:

- A alocação pode ocorrer primeiro quando há restrição de recursos. Dessa forma, a ferramenta otimiza a latência e o *throughput* (isto é, a quantidade de dados processados por unidade de tempo) do circuito a partir da quantidade de recursos disponível. É mais usado ao programar *chips* com poucas LUTs.

- Em contrapartida, o escalonamento pode tomar lugar antes da alocação quando há restrição de tempo. Assim, o algoritmo de síntese tenta otimizar a quantidade de recursos alocados e área utilizada dado o tempo máximo de cada operação. Essa estratégia pode se fazer mais útil em aplicações críticas, como FPGAs automotivos.

- A execução das três fases podem ocorrer de forma concorrente e intercomunicativa, de forma que os três processos se otimizem mutuamente. Apesar desse ser o modelo ideal, ele cria um modelo complexo demais, que acaba não sendo possível de se aplicar no processo de síntese de alto nível em exemplos realistas.

Em geral, aplicações com restrições diferentes exigem ordens de execução diferentes. Restrições de recurso (*e.g.* área de implementação, quantidade de unidades funcionais etc) rodam primeiro a alocação, para estabelecer o máximo de recursos e área que o circuito poderá utilizar e, a partir disso, otimizar sua geração nos outros passos. Por outro lado, restrições de tempo exigem o uso prévio do escalonador para estabelecer a latência máxima do processamento dos dados e, em seguida, ocorrem as otimizações possíveis em cima desse primeiro resultado.

Um exemplo de figura está na figura 4.1.

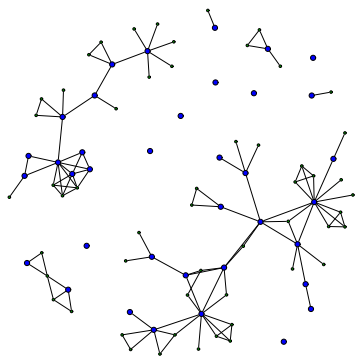

Figura 3.1: *Exemplo de uma figura.*

Capítulo 4

LegUp High-Level Synthesis

LegUp é um arcabouço *open-source* de síntese de alto nível desenvolvido na Universidade de Toronto (Canadá). Sua síntese converte códigos em C para Verilog e usa algumas ferramentas comerciais como o Quartus Prime II, da Intel, e o Tiger MIPS Processor, da Universidade de Cambridge (Reino Unido).

Atualmente em sua versão 4.0, apresenta uma arquitetura que permite alterações em seus algoritmos de forma relativamente simples, devido à sua modularização. Como sua arquitetura usa uma compilação em escopo de funções para implementação em *hardware* - isto é, ele usa funções como unidade básica para síntese de *hardware* -, é possível, por exemplo, especificar funções específicas para aceleração em *hardware* enquanto o resto do programa é executado em *software*; tal técnica é chamada de "fluxo híbrido"pelos criadores da ferramenta e é explicada melhor na seção **??**.

4.1 Fluxo de execução

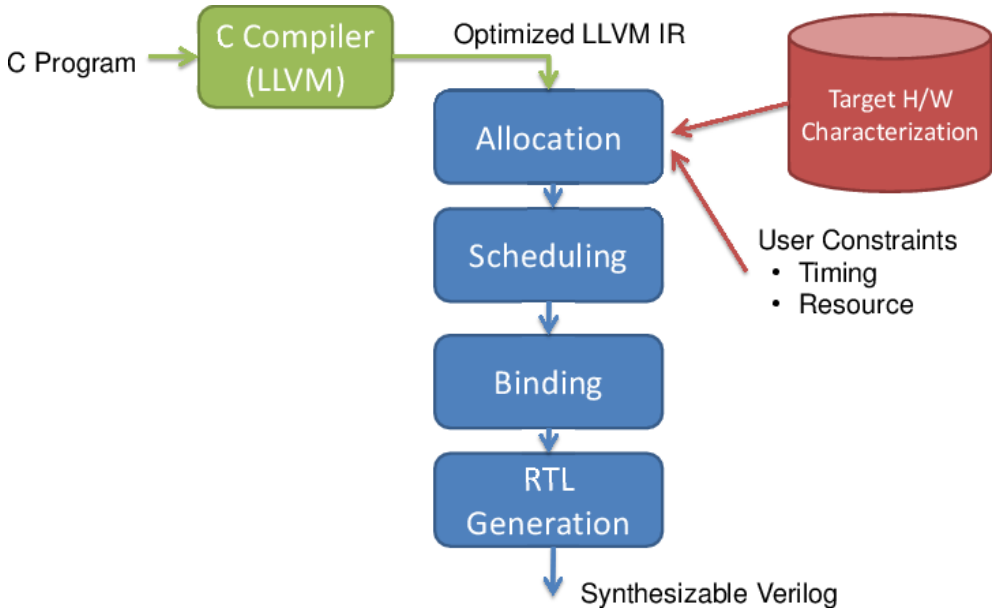

Figura 4.1: *Fluxo de execução do LegUp.*

A figura acima representa o fluxo geral do arcabouço. A entrada da ferramenta é um programa desenvolvido em C puro, que é compilado, otimizado e transformado em sua representação intermediária (IR) pela LLVM (vide apêndice A). Em seguida, na fase de

alocação, o LegUp usa os dados sobre o *hardware* no qual queremos implementar o algoritmo para alocar os recursos disponíveis no *chip*, tais como blocos de memória e unidades lógicas. Na etapa de escalonamento, as instruções da IR são mapeadas do grafo de controle e fluxo de dados para uma máquina de estados finita, onde cada estado é designado para um ciclo de *clock* específico. Depois esse mapeamento, o algoritmo de síntese atribui, a cada estado da máquina de estados, os recursos do *chip* necessários para a execução de suas instruções. Com as operações e recursos definidos, o arcabouço gera o RTL equivalente ao algoritmo e, por fim, o usa para criar um arquivo de descrição de *hardware*, escrito em Verilog.

Dada a forma como o arcabouço foi construído, isto é, na linguagem C++, utilizando o paradigma de programação orientada a objetos para modularizar o código, as etapas da síntese de alto nível feitas sobre o código compilado são implementadas em classes separadas, uma para cada etapa.

4.1.1 Fluxos de transformação

Apesar da existência do fluxo geral de funcionamento do LegUp, a ferramenta define três fluxos distintos chamados aqui de *fluxos de transformação*. Cada um deles transforma o programa de entrada em um tipo de circuito diferente, cada qual apresentando suas vantagens e desvantagens. Os fluxos implementados são o de puro *hardware*, puro *software*, e híbrido. O fluxo utilizado neste trabalho foi o de puro *hardware*, com o intuito de aproximar cientistas da computação à compreensão do processo de pesquisa e desenvolvimento em *hardware*.

Puro hardware

Neste fluxo, todo o programa de entrada do LegUp é transformado em *hardware*. Cada função do código é mapeada em um módulo Verilog que, ao ser compilado para o *chip* FPGA, funciona de forma paralela. Devido à paralelização inerente aos componentes de *hardware*, o controle do algoritmo é feito em um módulo Verilog chamado `main`, que descreve e controla a execução da máquina de estados finita que modela o algoritmo.

A maior vantagem desse fluxo é a velocidade de execução do algoritmo, que chega a ser 8 vezes maior (**colocar referencia**). Porém, ele não permite a implementação de técnicas importantes como recursão ou alocação dinâmica de memória.

Puro software

Neste fluxo, todo o programa de entrada do LegUp é transformado em *software*. Um processador *soft* (*softprocessor*) é instanciado pelo arcabouço, junto dos dados da aplicação, como instruções a serem executadas. Após a compilação da descrição de *hardware* na placa FPGA, o processador é executado no tecido FPGA como um processador comum. O processador usado pelo LegUp é descrito na subseção 4.1.2.

Utilizar esse fluxo dá a oportunidade de uso de técnicas importantes de programação, como recursão e alocação dinâmica de memória, ambas inviáveis via *hardware* puro. Além disso, ao executar um processador de forma isolada, o único processo existente para utilizá-lo é o da aplicação da FPGA, resultando em um menor *overhead* de troca de processos por parte de um sistema operacional. Entretanto, devido à frequência de *clock* de um *chip* FPGA ser da ordem de 10 vezes menor que o de um processador médio de um computador pessoal atual (**colocar referencia**), mesmo com a exclusividade de acesso do processo ao processador, a velocidade de execução pode ser muito inferior a um sistema embarcado com processador de uso geral implementado em um ASIC.

Fluxo híbrido

No fluxo híbrido, o programa de entrada é compilado de forma semelhante à feita no fluxo de puro *software*. A diferença principal é o fato de que o usuário pode definir marcações no código para dizer quais funções devem ser aceleradas por *hardware*, gerando um acelerador a ser usado na chamada da função especificada. Assim, chamadas dela no código de entrada são substituídas por *funções embrulhadas* (i.e. *wrapper functions*), que enviam um sinal para o acelerador executar o processamento de dados representados pela função. Nesse cenário, o processador tem duas opções quanto a seu funcionamento durante tal chamada: continuar executando o código da aplicação enquanto continuamente verifica se o acelerador terminou sua execução, ou esperar o acelerador terminar seu processamento e então, retomar a execução da aplicação. No LegUp, a segunda opção foi adotada na implementação da ferramenta.

O fluxo híbrido permite a aceleração de funções computacionalmente pesadas enquanto ainda dá abertura para o uso das técnicas de programação proibidas no fluxo de *hardware*. Porém, sua velocidade de processamento ainda é consideravelmente inferior ao do algoritmo totalmente implementado em *hardware*.

4.1.2 Compilação

O código usado como entrada do arcabouço deve ser escrito em C e possui limitações para certos fluxos. A versão gratuita mais recente da ferramenta não suporta implementações de recursão ou alocação dinâmica de memória; apesar disso, o LegUp consegue sintetizar estruturas, controles de fluxo, aritmética de inteiros, manipulação de ponteiros (inclusive ponteiros de funções), dentre outras características da linguagem.

A compilação do código é feita no *front-end* da LLVM usando o Clang 3.5, um compliador da linguagem C pertencente ao projeto LLVM, e cria um arquivo de *bytecode* contendo a LLVM IR correspondente ao programa de entrada. Algumas funções nativas da linguagem que lidam com o manejo da memória (como `memset` e `memcopy`, da biblioteca `string.h`) são compiladas pelo Clang em funções já implementadas pela LLVM, chamadas *funções intrínsecas*. Para contornar a situação, passes do otimizador são executados no código para substituir as funções intrínsecas para funções implementadas manualmente pelo arcabouço, gerando um *bytecode* composto da LLVM IR pura, sem funções intrínsecas.

Processador

O processador *soft* utilizado pelo arcabouço é o Tiger MIPS Soft Processor (https://www.cl.cam.ac.uk/t
um processador que pode ser implementado usando síntese lógica, isto é, seu comportamento pode ser descrito por uma linguagem de descrição de *hardware* (e.g. Verilog HDL) e então convertido em um *design* de *hardware*. Possuindo um tamanho de palavra (*word size*) de 32 bits, ele é usado na elaboração do circuito apenas nos fluxos híbrido e puro *software*, onde há a necessidade de um módulo central que controle o funcionamento do circuito. A vantagem de se usar o Tiger é seu código aberto e sua arquitetura RISC, que permitem a adição de novas instruções no processador, e de maneira menos complexa que a arquitetura CISC.

A possibilidade de modificação do processador do circuito é a característica chave do processo de autoavaliação que o LegUp realiza em seu fluxo de execução. Ao adicionar instrumentações para observar a execução do programa, é possível dizer quais instruções são mais utilizadas e por quais funções elas são mais chamadas. Isso permite uma análise extremamente precisa, uma vez que ela é feita a nível de instrução. Isso dá oportunidade

ao usuário de verificar as instruções resultantes da compilação do código em C e otimizá-las manualmente, de acordo com suas necessidades.

Apesar da utilização do Tiger, que é um *softprocessor*, a versão mais recente do LegUp open-source também dá suporte aos processadores *hard* da Altera e da Xilinx. Um processador *hard* não pode ser descrito por uma HDL e, por isso, seu design é construído de forma rígida no *chip*, como propriedade intelectual. O motivo dessa impossibilidade em descrevê-lo por uma HDL é pelo fato de que um processador *hard* tem sua construção especificada a nível de transístor, resultando em uma arquitetura muito específica para ser precisamente descrita por uma descrição de *hardware*. Apesar de afetar a flexibilidade de customização do processador, o uso desses tipos de processador aumenta a eficiência do FPGA em termos de energia, latência e área.

Nota: a partir da versão 5.0, o LegUp tornou-se comercial. Veja mais aqui(https://www.legupcomputi

4.1.3 Alocação de recursos

Essa etapa é feita pela classe `Allocation` da ferramenta, e usa *scripts* TCL para efetuar a alocação dos recursos presentes no *chip* FPGA. Um desses *scripts* contém a especificação do dispositivo, opções de síntese de alto nível e restrições de tempo; outro contém as restrições de área e latência de operações. Todas essas informações são armazenadas em uma instância da classe para que os estágios seguintes da síntese possam usá-las.

4.1.4 Escalonamento

Cada função do código de entrada é transformado em uma função na LLVM IR durante a compilação. O escalonador do LegUp transforma cada uma dessas funções em um objeto da classe `FiniteStateMachine`, que representa a máquina de estados finita daquela função. Cada objeto desses contém objetos da classe `State` que guarda cada estado da máquina de estados; este, por sua vez, contém instâncias da classe *InstructionNodes* que guarda informações sobre as instruções a serem executadas no estado correspondente, tais como suas latências.

As instâncias de `InstructionNodes` são criadas por uma classe chamada `SchedulerDAG`, responsável por ler cada instrução do programa e calcular as dependências de memória e de dados entre elas e, depois, inserir nas instâncias tais cálculos. Depois do cálculo de dependências, o escalonador mapeia cada `InstructionNodes` para seus respectivos estados através da classe `SchedulerMapping`.

A estratégia adotada pelo escalonador é baseada na formulação matemática das dependências como um problema de otimização linear, chamado *sistema de restrições de diferença*, como descrito em (J. Cong and Z. Zhang, ?An Efficient and Versatile Scheduling Algorithm Based On SDC Formulation,? Proceedings of the 2006 Design Automation Conference, San Francisco, CA, pp. 433-438, July 2006.). Nessa formulação, o programa linear contém restrições da forma

$$x_1 - x_2 \; REL \; y \tag{4.1}$$

onde

$$REL \in \{\leq, \geq, =\} \tag{4.2}$$

No processo de criação do sistema linear, o arcabouço mapeia as operações para serem feitas o mais cedo possível, dadas as dependências entre elas. Tal estratégia, chamada

as-soon-as-possible scheduling ou *ASAP scheduling*, pode ser trocada para outra, oposta, chamada *as-late-as-possible scheduling* ou *ALAP scheduling*. Finalmente, após a modelagem do programa linear com as operações e suas dependências, o sistema é resolvido utilizando-se a biblioteca *open source* `lpsolve` (http://lpsolve.sourceforge.net/).

Período de *clock* do *chip* utilizado, estratégia de escalonamento (*ALAP* ou *ASAP*), dentre outras informações importantes para o processo de escalonamento são encontradas em arquivos TCL pelos diretórios da ferramenta.

4.1.5 Emparelhamento

4.1.6 Geração do RTL

Capítulo 5

Conclusões

Texto texto[1].

[1]Exemplo de referência para página Web: www.vision.ime.usp.br/~jmena/stuff/tese-exemplo

Apêndice A

Projeto LLVM

LLVM (antigo acrônimo para "Low-level virtual machine") é um projeto de código aberto que dispobiliza ferramentas de compilação e otimização para diversas linguagens. Tais ferramentas conseguem compilar códigos de diferentes linguagens e otimizá-los em tempo de compilação, provido de um *front-end* e um *back-end* do usuário. Por *front-end* entende-se um *parser* e um *lexer* da linguagem de programação a qual se deseja compilar, enquanto que por *back-end* entende-se uma lógica de transformação do código próprio da LLVM em código de máquina. Um exemplo de uma ferramenta famosa pertencente ao projeto LLVM é o (Clang)(http://clang.llvm.org/), um compilador de C/C++/Objective-C alternativo ao GCC, que (pode apresentar perfomances superiores a este)(http://clang.llvm.org/features.html#performance).

Neste apêndice, serão apontadas características do projeto de forma direcionada ao entendimento do LegUp, descrito no capítulo **??**.

A.1 Estrutura

Figura A.1: *Estrutura básica de um compilador.*

A arquitetura mais utilizada na construção de um compilador é a chamada *arquitetura trifásica*, apresentando um *front-end*, um otimizador de código, e um *back-end*, como mostra a figura acima. O *front-end* é responsável pela transformação do arquivo de entrada em algum tipo de representação que permita sua leitura e otimização como, por exemplo, os *bytecodes* da linguagem Java. O otimizador recebe uma representação de um programa e realiza otimizações no código, que podem diminuir seu tempo de execução e/ou reduzir a quantidade de memória utilizada em sua execução. Por fim, o *back-end* converte o código otimizado na representação final desejada (também chamada de "*target*"ou "alvo"), que pode consistir em diversas representações, tais como um arquivo de texto simples que descreve o programa, ou um arquivo binário compatível com processadores da arquitetura x86. A LLVM também adota esse tipo de arquitetura, como visto na figura A.2.

A principal vantagem de se adotar esse tipo de estrutura é a modularização do sistema, resultando na possibilidade de se reutilizar partes do sistema para novas aplicações. Por exemplo: se existir uma aplicação cujo *front-end* recebe um código em Python, com um

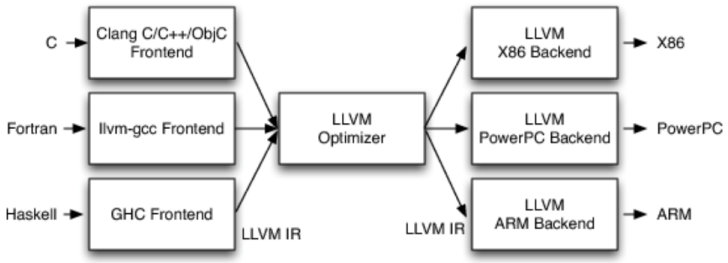

Figura A.2: *Abstração da implementação do Projeto LLVM.*

otimizador do código gerado pelo *front-end*, e um *back-end* que gera o código equivalente em Java, e houvesse a necessidade de mudar o alvo de Java para Haskell, não seria necessário reescrever todo o sistema apenas para mudar o *back-end*: bastaria mudar apenas a geração do código em Haskell, sem precisar repensar o resto do código.

A LLVM, além de adotar essa arquitetura, também apresenta uma forte modularização em seu código, através da orientação a objetos da linguagem C++. Isso porque aplicações como o GCC, ainda que sigam a arquitetura trifásica, possuem módulos altamente acoplados, tal que o desenvolvimento do *back-end* necessita do conhecimento do *front-end* e vice-versa. Esses tipos de aplicações são chamadas de *monolíticas*, ou seja, aplicações que na prática, são muito acopladas, com dependências difíceis de serem desfeitas sem alterar partes críticas e variadas do sistema.

A.2 Representação intermediária

As implementações e detalhes de ambos *front-end* e *back-end* dependem muito da aplicação para qual a LLVM está sendo usada. O *front-end* pode consistir de um *parser* e *lexer* de uma linguagem totalmente nova, cuja sintaxe siga um padrão bem diferente das linguagens já existentes, ou até um novo paradigma. O *back-end*, por sua vez, pode transformar o código em instruções ou outros códigos de outras linguagens, como [Scratch](https://scratch.mit.edu/about), destinadas a robôs feitos de peças Lego, ou até um texto simples que contém o número de instruções do programa compilado em cada uma das arquiteturas de hardware existentes. Como as possibilidades são muitas, o projeto adotou um tipo de representação de código utilizado em sua arquitetura, a chamada *representação intermediária da LLVM*, mais conhecida como *LLVM IR* ("*LLVM intermediate representation*"). Esta é enviada do *front-end* ao otimizador, onde é modificada de acordo com as regras descritas pelos desenvolvedores da aplicação e, depois, encaminhada para o *back-end* construir a saída apropriada para o alvo da aplicação. Um exemplo da LLVM IR pode ser visto abaixo.

```
define i32 @add1(i32 %a, i32 %b) {
entry:
  %tmp1 = add i32 %a, %b
  ret i32 %tmp1
}
```

O código acima é a *representação textual* da LLVM IR, uma vez que ela também pode ser serializada em *bitcode*, isto é, ter uma representação binária. O código define uma função chamada “‘add1“‘, que recebe dois inteiros “‘a“‘ e “‘b“‘ e retorna a soma deles. Como é possível perceber para quem já estudou ou viu códigos de alguma linguagem de montagem,

a LLVM IR se assemelha a esse tipo de linguagem, de uma arquitetura RISC. O equivalente da função, em C, seria:

```
unsigned int add1(unsigned int a, unsigned int b) {
    unsigned int tmp1 = a + b;
    return tmp1;
}
```

O uso dessa representação intermediária facilita o desenvolvimento de uma aplicação ao padronizar a saída do *front-end* e a entrada do *back-end*, bem como partes do otimizador. Assim, ao criar um novo *front-end* para a LLVM, por exemplo, um programador deve saber apenas as características da entrada e da LLVM IR. Como o otimizador e o *back-end* utilizam a LLVM IR de forma independente, não é necessário saber sobre eles para a execução de seu trabalho.

A.3 LLVM Pass Framework

No meio do processo de compilação, e considerando a arquitetura trifásica, encontra-se o otimizador do código. Ele é responsável por realizar modificações que melhorem, por exemplo, o tempo de execução do programa e o uso de espaço de memória do computador. No caso da LLVM, o otimizador recebe um código descrito pela LLVM IR e altera as instruções ao reconhecer determinados padrões. Por exemplo, se houver uma instrução onde há a subtração de um número inteiro por ele mesmo é atribuída a uma variável:

```
...
%tmp1 = sub i32 %a, %a
...
```

É possível, ao invés disso, atribuir 0 à variável:

```
%tmp1 = i32  0
```

Ou seja, reconhecendo um padrão na instrução (e.g. subtração de um inteiro por ele mesmo), substitui-se a instrução por outra mais eficiente (e.g. atribuir 0 à variável).

O mecanismo empregado na LLVM para realizar essas otimizações são os chamados *passes*, do arcabouço *LLVM Pass Framework*, pertencente ao projeto. Em termos práticos, os passes são etapas, possivelmente independentes entre si, pelas quais o código (ou parte dele) passa por uma análise e busca de padrões desejados em instruções e suas possíveis alterações; em termos técnicos, os passes são classes derivadas da superclasse `Pass` direta ou indiretamente, que indicam o escopo mínimo pelo qual o passe é responsável (e.g. escopo global, de função, de bloco básico, de *loop*) e que implementam interfaces usados pelo arcabouço para realizar as otimizações. Cada passe é, assim, responsável por identificar padrões de instrução dentro do seu escopo e otimizar o padrão observado. A alteração retratada acima, onde temos a subtração de um inteiro por ele mesmo trocada pela atribuição da variável pelo valor 0, poderia ser colocada dentro de um passe junto de outras otimizações com respeito à aritmética de inteiros, como transformar $x - 0$ em x.

Referências Bibliográficas

Tufte(2001) Edward Tufte. *The Visual Display of Quantitative Information.* Graphics Pr, 2nd edição. Citado na pág. 1

Wazlawick(2009) Raul S. Wazlawick. *Metodologia de Pesquisa em Ciencia da Computação.* Campus, primeira edição. Citado na pág. 1

Zobel(2004) Justin Zobel. *Writing for Computer Science: The art of effective communication.* Springer, segunda edição. Citado na pág. 1