

MS Technical Paper: Placement Algorithms for Heterogeneous FPGAs

Brian B Cheng

Rutgers University Department of Electrical and Computer Engineering

1 Keywords

- FPGA, EDA, Placement, Simulated Annealing, Optimization, RapidWright

2 Abstract

Placement remains one of the most computationally demanding stages of FPGA computer-aided design (CAD) flows, directly influencing routability, timing, and design productivity. This paper presents a modular placement framework targeting Xilinx 7-Series devices, implemented using the RapidWright API. The flow first clusters multi-cell structures (such as CARRY4 chains, DSP48E1 cascades, and LUT-FF pairs) into prepacked groups, then maps them to physical sites before applying a simulated annealing (SA) placer. The SA engine minimizes half-perimeter wirelength (HPWL) using a geometric cooling schedule, with support for both random and directed moves. The result is a legal, wirelength-optimized placement that forms a foundation for experimenting with alternative cost functions and analytical or hybrid placement strategies.

3 Introduction

Field-Programmable Gate Arrays (FPGAs) have witnessed rapid growth in capacity and versatility, driving significant advances in computer-aided design (CAD) and electronic design automation (EDA) methodologies. Since the early-to-mid 2000s, the stagnation of single-processor performance relative to the rapid increase in integrated circuit sizes has led to a design productivity gap, where the computational effort for designing complex chips continues to rise. FPGA CAD flows mainly encompass synthesis, placement, and routing; all of which are NP-hard problems, of which placement is one of the most time-consuming processes. Inefficient placement strategy not only extends design times from hours to days, thereby elevating cost and reducing engineering productivity, but also limits the broader adoption of FPGAs by software engineers who expect compile times akin to those of software compilers like `gcc`.

For these reasons, FPGA placement remains a critical research effort even today. In this paper, we study and implement established placement methods. To do this, we use the RapidWright API, which is a semi-open-source research effort from AMD/Xilinx that enables custom solutions to FPGA design implementations and design tools that are not offered by their industry-standard FPGA environment, Vivado. We implement multiple variations of

simulated annealing placers for Xilinx's 7-series FPGAs, with an emphasis on minimizing total wirelength while mitigating runtime. Our implementation is organized into three consecutive substages. The **prepacking** stage involves traversing a raw EDIF netlist to identify recurring cell patterns—such as CARRY chains, DSP cascades, and LUT-FF pairs—that are critical for efficient mapping and legalization. In the subsequent **packing** stage, these identified patterns, along with any remaining loose cells, are consolidated into **SiteInst** objects that encapsulate the FPGA's discrete resource constraints and architectural nuances. Finally, the **placement** stage employs a simulated annealing (SA) algorithm to optimally assign **SiteInst** objects to physical **Sites**, aiming to minimize total wirelength while the constraints of the 7-series architecture.

Simulated annealing iteratively swaps placement objects guided by a cost function that decides which swaps should be accepted or rejected. Hill climbing is permitted by occasionally accepting moves that increase cost, in hope that such swaps may later lead to a better final solution. SA remains a popular approach in FPGA placement research due to its simplicity and robustness in handling the discrete architectural constraints of FPGA devices. While SA yields surprisingly good results given relatively simple rules, it is ultimately a heuristic approach that explores the vast placement space by making random moves. Most of these moves will be rejected, meaning that SA must run many iterations, usually hundreds to thousands, to arrive at a desirable solution.

In the ASIC domain, where placers must handle designs with millions of cells, the SA approach has largely been abandoned in favor of analytical techniques, owing to SA's runtime and poor scalability. Modern FPGA placers have also followed suit, as new legalization strategies allow FPGA placers to leverage traditionally ASIC placement algorithms and adapt them to the discrete constraints of FPGA architectures. While this paper does not present a working analytical placer, it will explore ways to build upon our existing infrastructure (prepacker and packer) to replace SA with AP.

The paper first begins by elaborating on general FPGA architecture and then specifically the Xilinx 7-Series architecture. Then, the paper will elaborate on the FPGA design flow, then the role that the RapidWright API plays in the design flow. We explain in detail each of these concepts for a broader audience as they provide much needed context for FPGA placement algorithms as a concept. However, readers who are already familiar with these concepts can skip directly to the RapidWright API section 7 or to the Simulated Annealing section .

4 FPGA Architecture History

Before any work can begin on an FPGA placer, it is necessary to understand both the objects being placed and the medium in which they are placed. Configurable logic devices have undergone significant evolution over the past four decades. We will briefly review the evolution of configurable logic architecture starting in the 1970s and quickly work our way up to modern day FPGA architecture.

PLA: The journey began with the Programmable Logic Array (PLA) in the early 1970s. The PLA implemented output logic using a programmable-OR and programmable-AND plane that formed a sum-of-products equation for each output through programmable fuses. Around the same time, the Programmable Array Logic (PAL) was introduced. The PAL simplified the PLA by fixing the OR gates, resulting in a fixed-OR, programmable-AND design, which sacrificed some logic flexibility to simplify its manufacture. Figure 1 shows one such PAL architecture.

CPLD: Later in the same decade came the Complex Programmable Logic Device (CPLD), which took the form of an array of Configurable Logic Blocks (CLBs). These CLBs were typically modified PAL blocks that included the PAL itself along with macrocells such as flip-flops, multiplexers, and tri-state buffers. The CPLD functioned as an array of PALs connected by a central programmable switch matrix and could be programmed using a hardware description language (HDL) like VHDL. Figure 2 shows one such CPLD architecture.

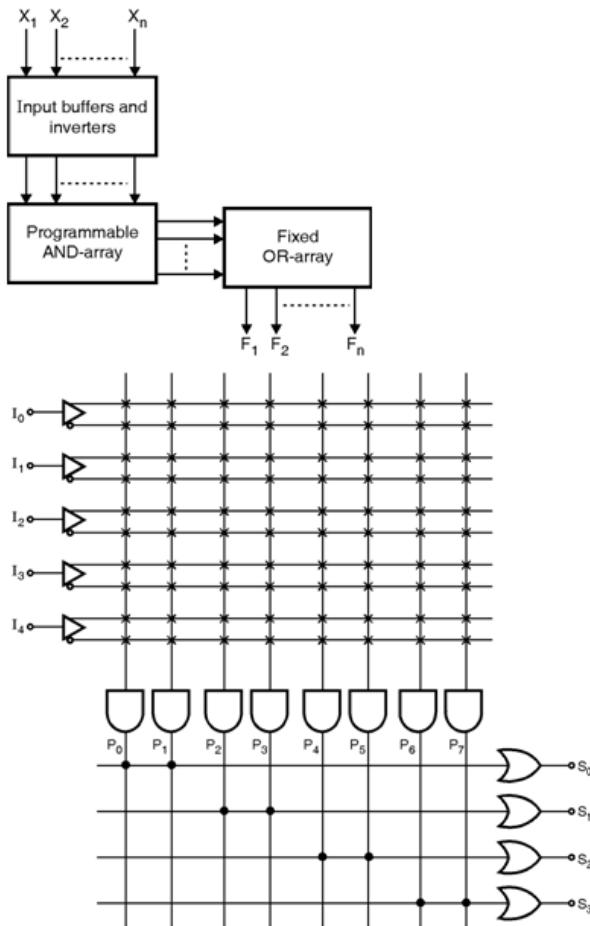


Figure 1: PAL architecture with 5 inputs, 8 programmable AND gates and 4 fixed OR gates

Homogeneous FPGA: The mid-1980s saw the introduction of homogeneous FPGAs, which were built as a grid of CLBs. Rather than using a central programmable switch matrix as in CPLDs, FPGAs adopted an island style architecture in which each CLB is surrounded on all sides by programmable routing resources, as shown in Figure 3. The first commercially viable FPGA, produced by Xilinx in 1984, featured 16 CLBs arranged in a 4x4 grid. As FPGA technology advanced, CLBs were redesigned to use lookup tables (LUTs) instead of PAL arrays for greater logic density. The capacity of an FPGA was often measured by how many logical elements or CLBs it offered, which grew from hundreds to thousands and now to hundreds of thousands of CLBs.

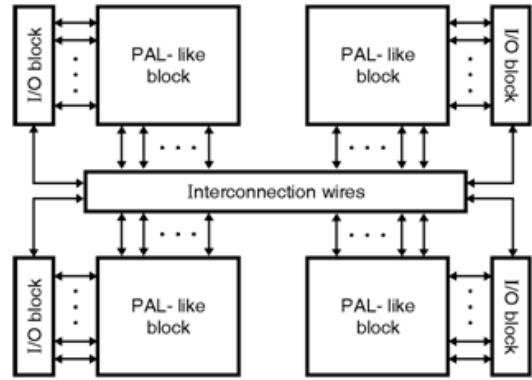


Figure 2: CPLD architecture with 4 CLBs (PAL-like blocks)

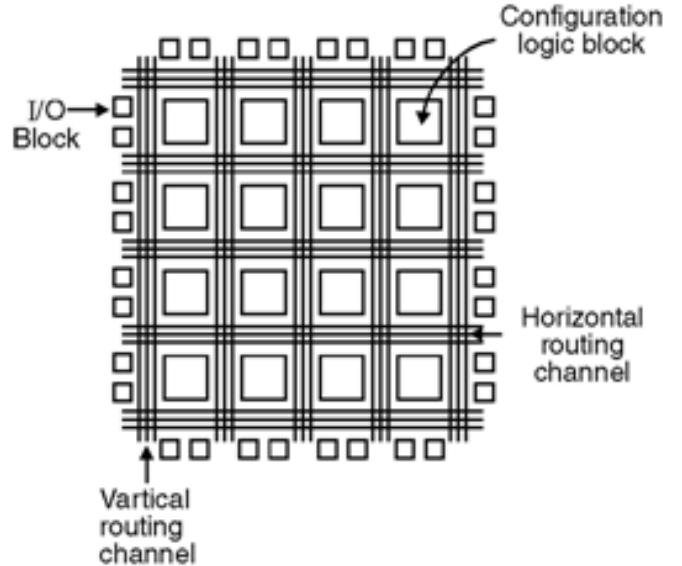


Figure 3: A homogeneous island-style FPGA architecture with 16 CLBs in a grid.

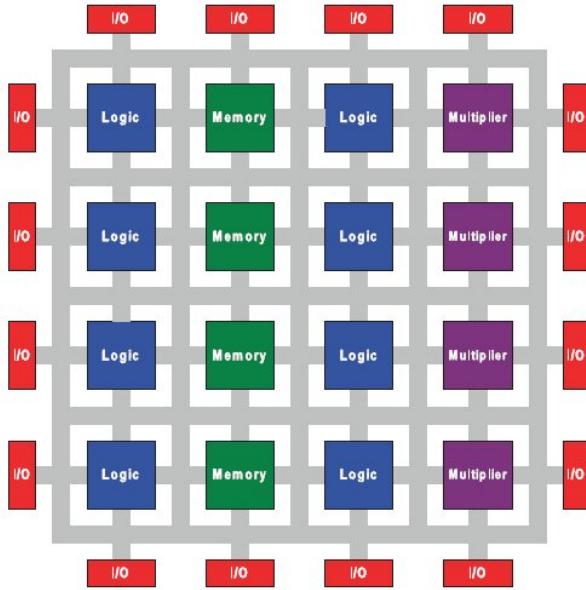


Figure 4: A heterogeneous island-style FPGA with a mix of CLBs and macrocells.

Heterogenous FPGA: This brings us to modern day FPGA architectures. To meet the needs of increasingly complex designs, FPGA vendors introduced heterogeneous FPGAs. In these devices, hard macros such as Block RAM (BRAM) and Digital Signal Processing (DSP) slices are integrated into the programmable logic fabric along with CLBs, like shown in Figure 4. This design enables the direct instantiation of common subsystems like memories and multipliers, without having to recreate them from scratch using CLBs. Major vendors such as Xilinx (AMD) and Altera (Intel) now employ heterogeneous island-style architectures in their devices. As designs be-

come increasingly large and complex, FPGAs meet the demand by becoming increasingly heterogenous, incorporating a wider variety of hard macros into the fabric.

5 Xilinx 7-Series Architecture

The Xilinx 7-Series devices, first introduced in 2010, follow a heterogeneous island-style architecture as discussed previously. Although the 7-Series was later superseded in 2013 by the UltraScale architecture, the 7-Series remains highly relevant due to its accessibility, wide availability, and compatibility with open-source tooling. Representative sub-families include Artix-7, Kintex-7, Virtex-7, and Zynq-7000, each designed with different performance and cost trade-offs but all follow the core 7-Series architecture.

Figure 5 illustrates a high-level view of the hierarchical organization of a 7-Series FPGA. At its lowest level, the device consists of a large array of atomic components called *Basic Elements of Logic* (**BELs**). These BELs encompass look-up tables (**LUTs**), flip-flops (**FFs**), block RAMs (**BRAMs**), DSP slices (**DSPs**), and the configurable interconnect fabric. They constitute the fundamental building blocks for implementing digital circuits on the FPGA.

To manage this complexity, Xilinx organizes these **BELs** into incrementally abstract structures. First, **BELs** are grouped into **Sites**. Each Site is embedded into a **Tile**, and Tiles are further arranged into **Clock Regions**. Note how the Tile arrangement is columnar such that column consists of only one Tile type. In some high-density devices, multiple Clock Regions may be consolidated into one or more **Super Logic Regions** (SLRs). However, for the scope of this paper, we focus on Xilinx 7-Series devices with only a single SLR.

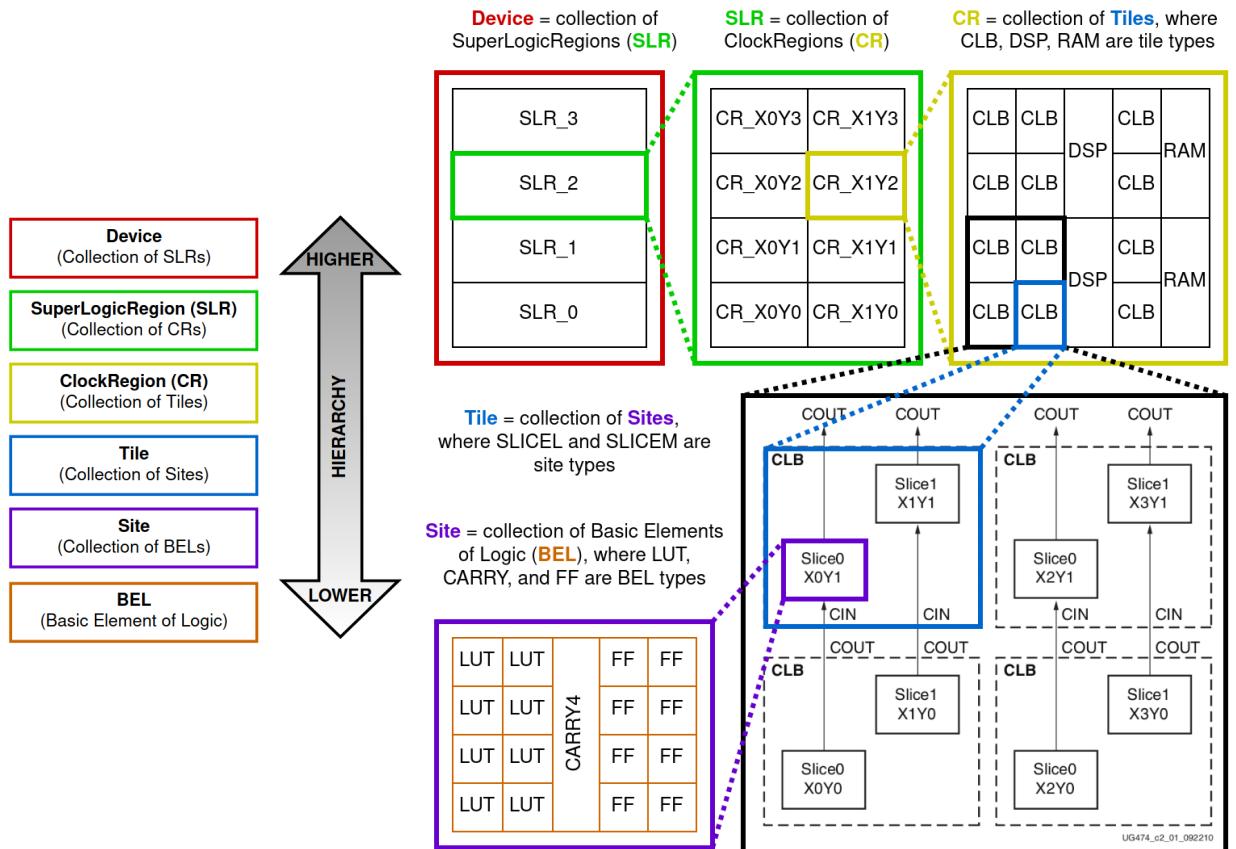


Figure 5: Architecture Hierarchy of a Xilinx FPGA

5.1 CLB SLICES

In the 7-Series architecture, the term *CLB* (Configurable Logic Block) refers to a *CLB Tile* that contains two *SLICE* Sites. Xilinx offers two variants of SLICE Sites: **SLICEL** and **SLICEM**.

- Each **SLICEL** has a set of BELs including eight LUTs, eight FFs, and one CARRY4 adder. The LUT BELs in a **SLICEL** can only host LUT Cells.
- The **SLICEM** includes all the features of a **SLICEL** but its LUT BELs can host both LUT Cells, which are asynchronous ROM elements, or **RAM32M** Cells, which are synchronous 32-deep RAM elements. These cells are also referred to as *Distributed RAM* in the Xilinx documentation. These cells can offer an alternative to the larger, more dedicated 18K-36K **RAMB18E1** cells when RAM resources are highly utilized or when the design inherently demands homogeneously distributed local memory.

In a typical 7-Series device, approximately 75% of the SLICE Sites are **SLICELs** and 25% are **SLICEMs**. A single CLB Tile can therefore host either two **SLICELs** or one **SLICEL** and one **SLICEM**. However, to simplify the problem space, we will only consider **SLICELs** for general logic and use the dedicated **RAMB18E1** cells to implement RAM elements.

The BELs in these SLICEs facilitate the bulk of the general programmability of the FPGA fabric. We will explain in detail the function and motivation behind these BELs.

LUTs Combinational logic is universal to all HDL designs. As their name suggests, a Look-Up Table (LUT) map an input value to an output value. LUTs facilitate combinational logic by acting as tiny asynchronously-accessed ROMs whose contents are fixed when the FPGA is programmed. For any boolean function, the synthesizer precalculates the boolean output to every possible input combination and stores the resulting truth table into a LUT's static memory. The inputs are then essentially treated as an address space that maps to a data value space in an asynchronous ROM. No explicit logic gates like NAND or XNOR are synthesized, contrary to what the name *Field Programmable Gate Array* might suggest.

In the 7-Series devices, one LUT can facilitate any 6-input boolean function, or two 5-input functions, as long as they share the same input signals. The LUT can also host two independent boolean functions of up to 3 inputs each, even when the inputs are not shared. Functions requiring more than six unique inputs are decomposed across multiple cascaded LUTs. Figure 6 shows an example of where a LUT is typically synthesized in a design entry.

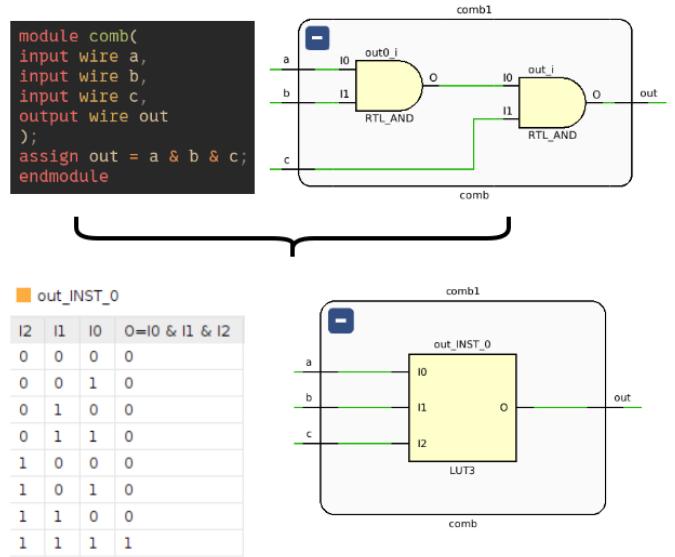


Figure 6: LUT synthesis from user design

FFs FFs are synthesized to facilitate synchronous event-driven signal assignment. For most Verilog users, this generally means signal assignments wrapped in `always @(posedge clk)` statements. Figure 7 shows an example of where a FF is typically synthesized. The cell primitive **FDRE** is a type of FF and belongs to a family of D Flip Flops (DFFs) with Clock Enable (CE).

- FDCE - DFF with CE and Asynchronous Clear
- FDPE - DFF with CE and Asynchronous Preset
- FDSE - DFF with CE and Synchronous Set
- FDRE - DFF with CE and Synchronous Reset

In a typical HDL design, the vast majority of FFs will be synthesized as FDREs with the occasional FDSE, as it is generally good practice to keep FPGA designs synchronous. A FF BEL may also host a **LATCH** Cell, however, since they are generally bad practice in FPGA design, we will not consider latches in this paper.

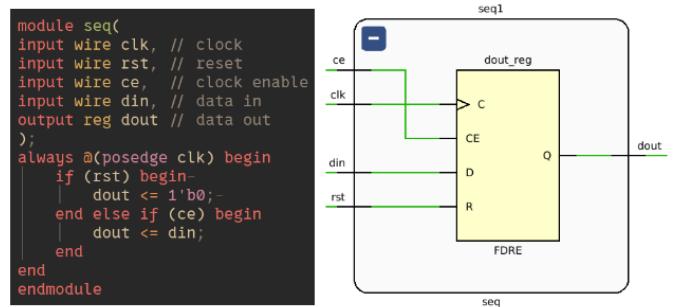


Figure 7: FF synthesis from user design

Up to eight (8) FFs can be placed within the same SLICE, but only if they all share a common Clock-Enable (CE) net and a common Set-Reset (SR) net. This is because the SLICE has only one CE pin and one SR pin to interface with general routing. The CE and SR signals from these pins are broadcast intra-Site to all FFs within.

LUT-FF Pairs FPGA designs are very often modelled as a collection of Finite State Machines (FSM) like shown in Figure 8. Many times a design will also use pipelining, either to model signal buffers or shift registers, or to split up large combinational logic blocks into time slices to meet timing constraints. These common design structures result in many consecutive sections of combinational logic feeding into a vector of registers. The synthesizer naturally synthesizes these structures as consecutive pairs of LUTs feeding into FFs as shown in Figure 9. Figure 10 shows an example of a synthesized LUT-FF Pair.

Shown in Figure 11 are two possible placements for a LUT-FF Pair on the physical device. On the right, the cells are placed across different Sites, thus the only way to route the net between the cells is through general inter-site routing. On the bottom left, the cells are placed within the same Site in the same lane, taking advantage of the intra-site routing without burdening the general router with additional inter-site routing. This is an important consideration to make while minimizing wirelength during placement.

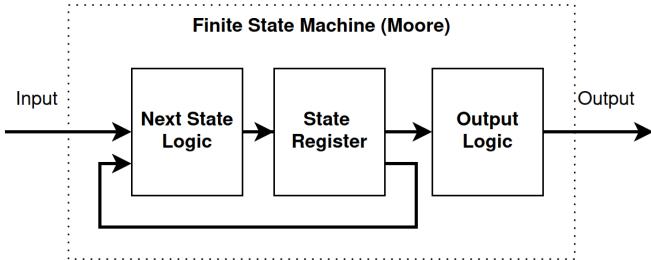


Figure 8: Finite state machine (Moore)

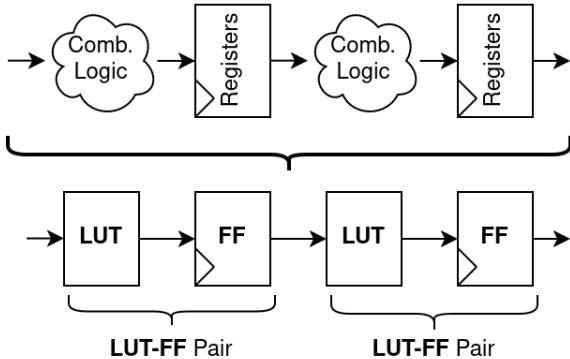


Figure 9: Pipelining synthesized as consecutive LUT-FF pairs

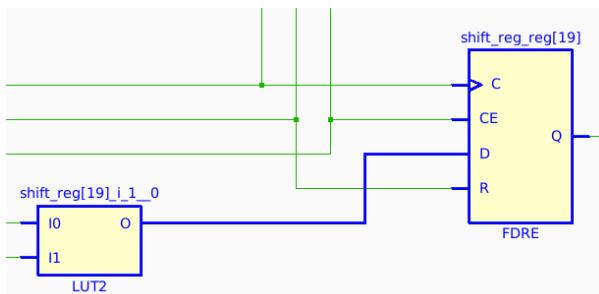


Figure 10: A synthesized LUT-FF Pair

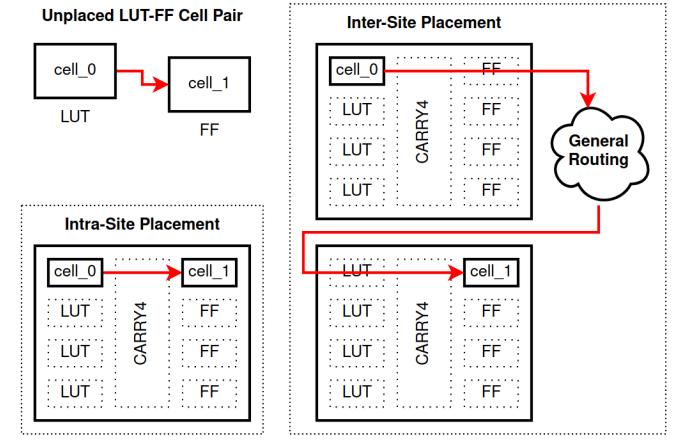


Figure 11: Intrasite vs Intersite LUT-FF Placement

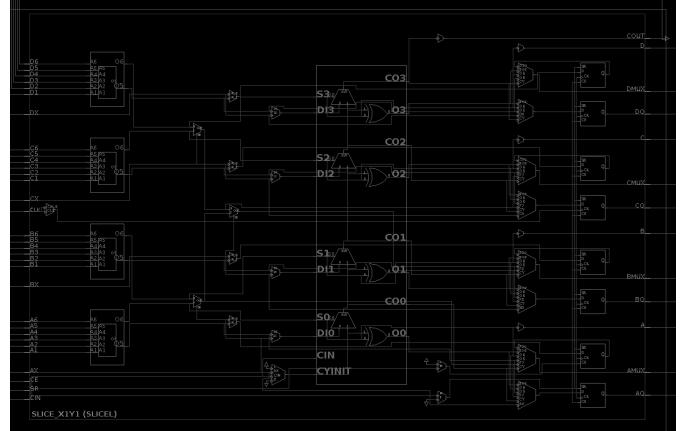


Figure 12: A SLICEL Site

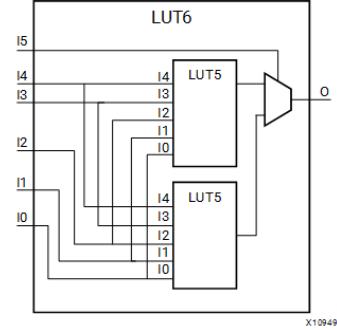


Figure 13: A LUT6 Cell is hosted using two LUT5 BELs.

A group of LUT-FF pairs may be placed in the same SLICE, with the constraint that the FFs must share the same Clock-Enable (CE) and Set-Reset (SR) nets. Clustering LUT-FF pairs like this can reduce the redundancy of having to route the same CE and SR nets to many SLICES by routing the nets to fewer SLICES and connecting them to the individual FFs within via intra-Site routing, and at the same time, packing greater logic density over a smaller area of the device. Recall that up to eight FFs may be placed in the same SLICE, thus theoretically, up to eight LUT-FF pairs can be placed in the same SLICE.

Utilizing all 8 LUT-FF lanes in a SLICE can help reduce device area utilization and minimize wirelength, but *too much* logic density in an area can contribute to general routing congestion. Furthermore, attempting to fill all 8 lanes in a SLICE requires meticulous adherence to conditional LUT constraints. Recall that a LUT can accommodate one 6-input boolean function or two 5-input boolean functions sharing the same inputs or any two 3-input or less boolean functions regardless of shared inputs.

A LUT6 Cell (a LUT with 6-inputs) will actually occupy two LUT-FF lanes in a SLICE, rendering one of the FF BELs in either lane ineligible for another LUT-FF pair. Figure 12 hints at this by depicting pairs of LUTs stacked ontop of one another. The LUTs individually can host up to a LUT5, but combined together can host one LUT6, like shown in Figure 13. To simplify the problem space, we will only fill up to four LUT-FF lanes in any given SLICE.

CARRY An FPGA design will also typically implement many adders, counters, subtractors, or comparators, all of which are based on binary addition. They are so ubiquitous that every that in the 7-Series architecture, every SLICE features a CARRY4 BEL – a 4-bit carry-lookahead (CLA) adder, a much better alternative to synthesizing adders via LUTs.

CARRY Chains These CARRY4 blocks can be chained across SLICES to implement wide adders efficiently. The CARRY4 BELs *must* be chained vertically consecutively across SLICES as the Carry-In (CI) and Carry-Out (CO) pins can only be routed this way. A CARRY4 cell may also directly connect to LUTs and FFs, and should be placed in the same Site whenever possible to minimize wirelength. Shown in figure 14 shows how a CARRY4 chain and associated LUTs and FFs can be placed across SLICES.

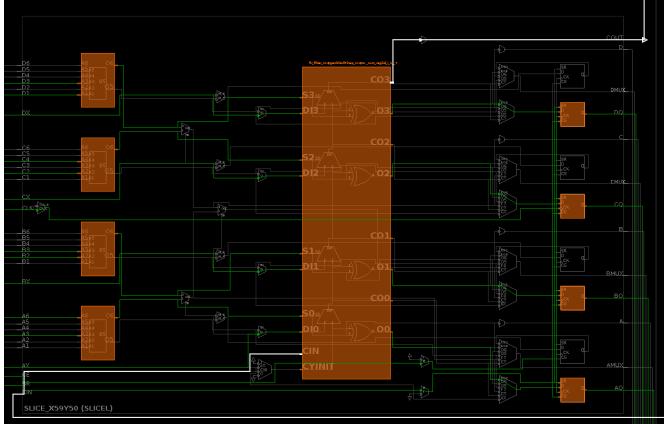


Figure 14: A SLICEL populated by a CARRY4 cell, 4 LUT cells, and 4 FF cells

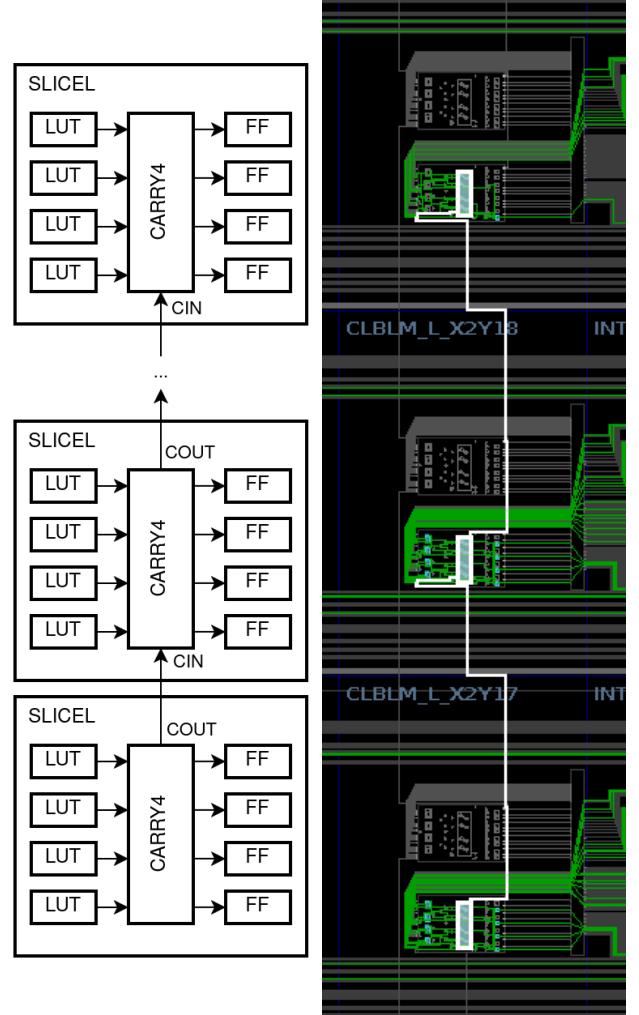


Figure 15: A CARRY4 chain of size 3 placed across 3 SLICES. **Left:** Simplified view, **Right:** As shown in the Vivado device viewer.

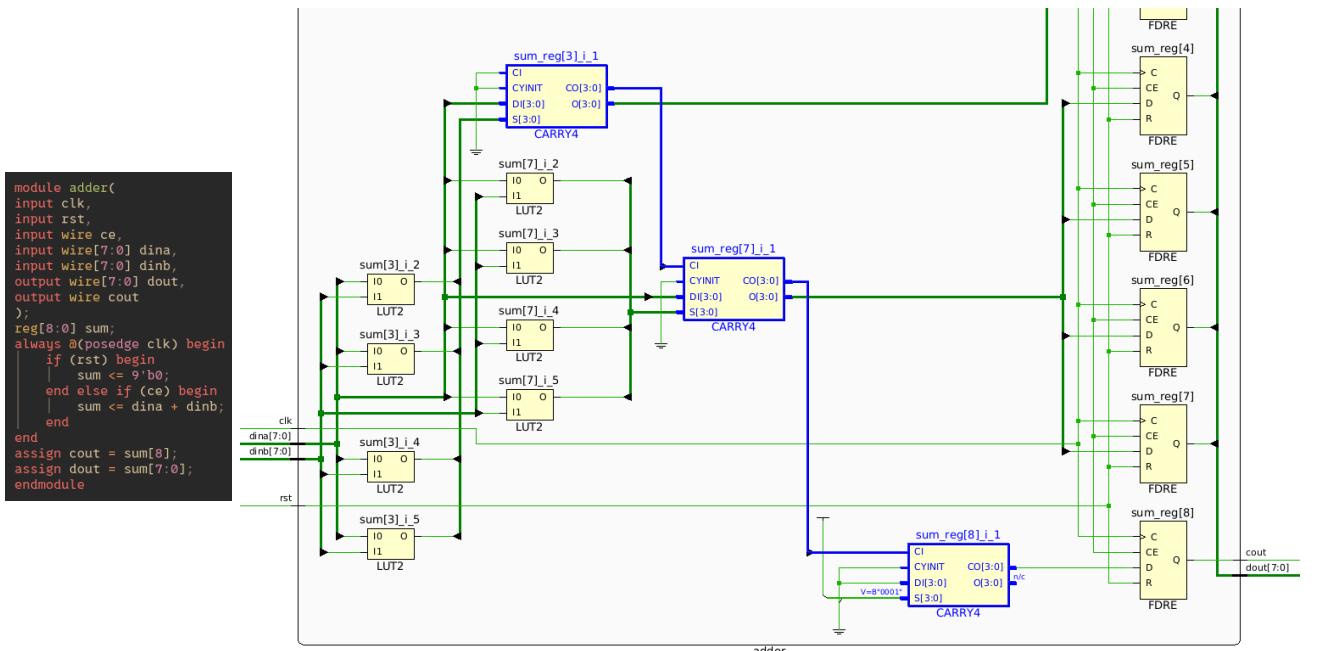


Figure 16: A CARRY4 chain of size 3 as shown in the Vivado netlist viewer

5.2 DSP Slices

DSPs FPGAs are often used as low latency Digital Signal Processing (DSP) accelerators. Common DSP subsystems like Finite Impulse Response (FIR) filters, Fast Fourier Transform (FFTs), and convolutional neural nets (CNNs) demand fast large scale multiply-accumulate (MAC) capabilities. The 7-Series architecture integrates DSP BELs into the logic fabric called DSP48E1 that can facilitate MAC efficiently. The architecture hierarchy for DSPs is simple compared to CLBs and SLICEs. A DSP48E1 Tile contains two DSP48E1 Sites, each containing one DSP48E1 BEL, which can host a DSP48E1 Cell.

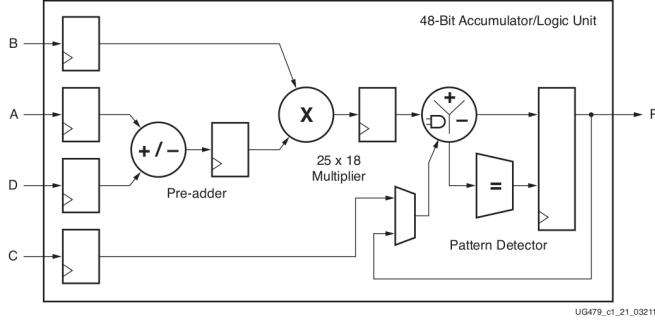


Figure 17: Basic DSP48E1 Slice Functionality

DSP Cascades Wider DSP functions are supported by cascading DSP48E1 slices in a DSP48E1 column. Much like CARRY4 chains, DSP48E1 cascades must necessarily be placed vertically consecutively across DSP48E1 Sites. They are connected by three busses: ACOUT to ACIN, BCOUT to BCIN, PCOUT to PCIN. These signal busses run directly between the vertical DSP48E1 slices without burdening the general routing resources. The ability to cascade this way provides a high-performance and low-power im-

plementation of digital signal processing systems.

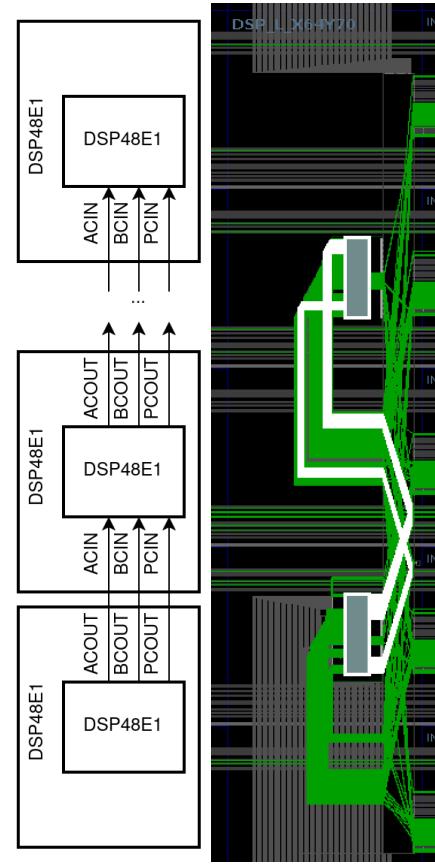


Figure 18: A DSP48E1 cascade of size 2 placed across 2 DSP48E1 Sites. **Left:** Simplified view, **Right:** As shown in the Vivado device viewer.

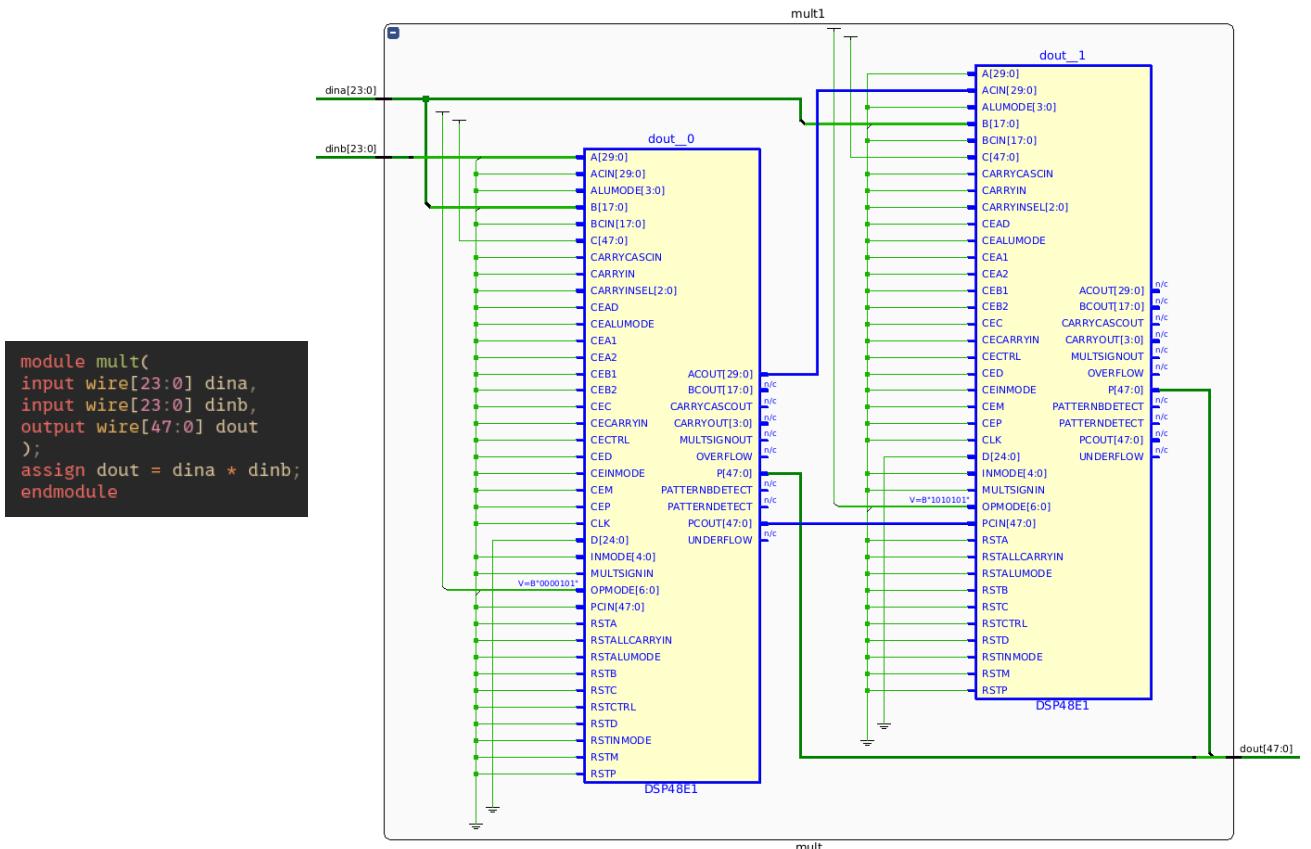


Figure 19: Simple multiplier synthesis from user design.

5.3 Block RAM

In addition to SLICEs and DSPs, the 7-Series also offers dedicated Block Random Access Memory (BRAM) BELs. These BRAMs come in two variants: **RAMB18E1** and **RAMB36E1**.

- **RAMB18E1** - Has a capacity of 18 Kilobits. It can be configured as single port RAM with dimensions ranging between (1-bit wide by 16K deep) to (18-bit wide by 1024 deep). It can also be configured as a (36-bit wide by 512 deep) true simple dual port RAM.
- **RAMB36E1** - Has a capacity of 36 Kilobits. It can be configured as single port RAM with dimensions ranging between (1-bit wide by 32K deep) to (36-bit wide by 1024 deep). It can also be configured as a (72-bit wide by 512 deep) simple dual port RAM.

One BRAM Tile contains one RAMB36E1 Site and two RAMB18E1 Sites. The RAMB36E1 Site contains one RAMB36E1 BEL, which can host one RAMB36E1 Cell. Likewise, the RAMB18E1 Site contains one RAMB18E1 BEL, which can host one RAMB18E1 Cell.

Like DSP cascades, BRAMs may also be cascaded together in a column to implement large memories efficiently, with some intermediate signals between them routed intra-Tile without burdening the general routing. However, unlike DSPs, large memories decomposed amongst multiple BRAMs can also be routed together through general routing. Furthermore, in most design scenarios, large memories will not utilize the intra-Tile signals.

To simplify the problem space, we will not constrain large BRAMs to be cascaded together in consecutive vertical Tiles. We can essentially treat RAMB18E1 and RAMB36E1 Cells as loose, minimally constrained single cells, in contrast to the highly constrained DSP48E1 cascades and CARRY4 chains.

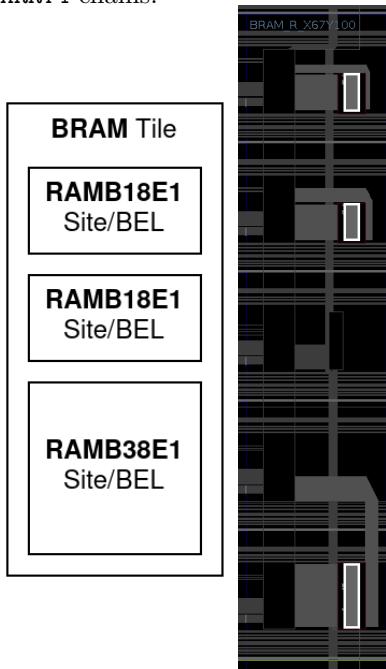


Figure 20: A BRAM Tile containing two RAMB18E1 Sites and one RAMB36E1 Site. **Left:** simplified view. **Right:** as seen in the device viewer, BELs highlighted in white.

```
module ram(
    input wire clk,
    input wire en,
    input wire we,
    input wire rst,
    input wire [9:0] addr,
    input wire [23:0] din,
    output wire [23:0] dout
);
reg [23:0] ram [1023:0];
reg [23:0] dout;
always @ (posedge clk)
begin
    if (en) begin
        if (we)
            ram[addr] <= din;
        if (rst)
            dout <= 0;
        else
            dout <= ram[addr];
    end
end
endmodule
```

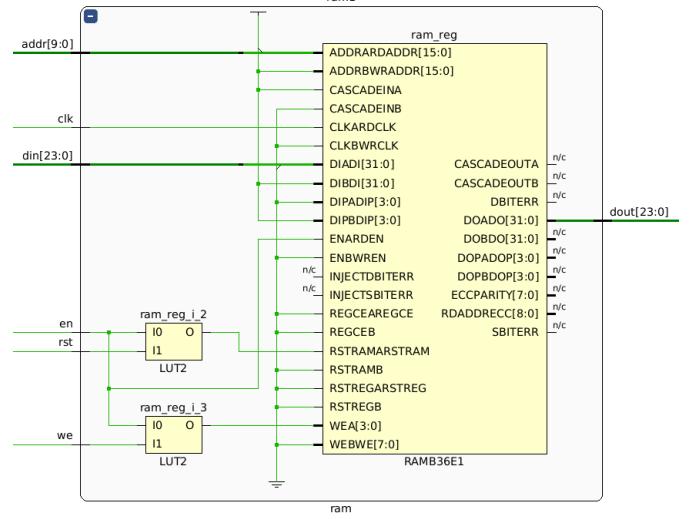


Figure 21: An example of BRAM synthesis (via inference)

5.4 Further Documentation

For more in-depth details about 7-Series FPGAs, refer to the official Xilinx user guides such as:

- *7 Series FPGAs Overview* (UG476)
- *7 Series FPGA Configurable Logic Block* (UG474)
- *7 Series Memory Resources* (UG473)
- *7 Series DSP48E1 Slice* (UG479)

This architectural context provides the necessary background for understanding how a placement algorithm should account for resource and placement constraints and optimize wirelength in modern FPGA architectures.

6 FPGA Design Flow and Toolchain

Modern FPGA designs require a sophisticated toolchain to bridge the gap between high-level hardware descriptions and the final bitstream used to configure the FPGA. Figure 22 illustrates a representative process that converts an abstract Hardware Description Language (HDL) design into a verified configuration file for a target device.

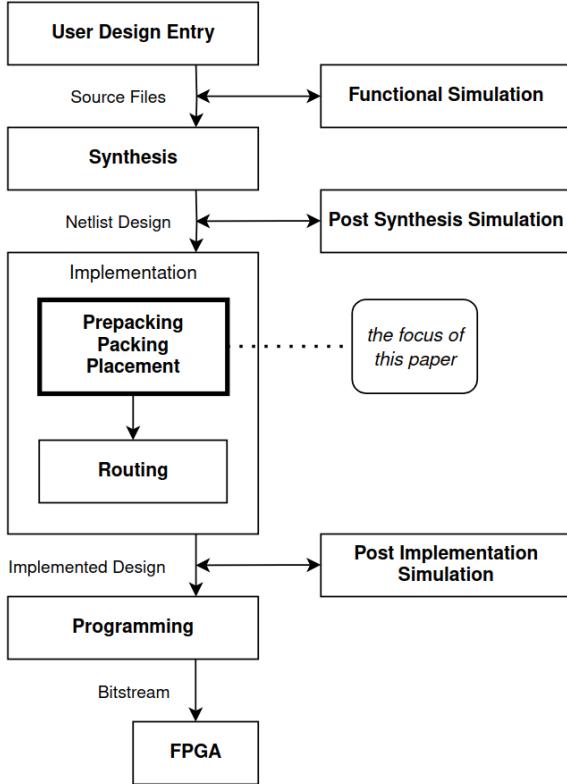


Figure 22: A typical FPGA design and verification workflow.

placement) to better meet timing and area objectives.

4. **Bitstream Generation:** After a design is fully placed, routed, and timing-closed, the toolchain produces a final *bitstream* that sets the configuration of every programmable element in the FPGA. This bitstream can then be loaded onto the device, either through vendor software or via a custom programming interface.
5. **Verification:** In parallel to the design flow, simulations and testbenches validate correctness of the user's design at multiple abstraction levels. Engineers may begin with behavioral simulations, then progress to post-synthesis simulations, and finally to post-implementation simulations that incorporate estimated routing delays. With each higher level of fidelity, computational requirements grow significantly due to increasing complexity and the need to analyze more variables over time. Ensuring correct functionality and meeting timing closure at the post-implementation stage is crucial before deploying the design to hardware. Given the importance of thorough verification, many established companies dedicate one verification engineer for every design engineer.

1. **Design Entry:** An engineer describes the intended functionality of the digital system using a hardware description language (HDL) such as Verilog or VHDL. During this phase, the coding style can vary (behavioral, structural, dataflow, etc.), but it generally aims to capture high-level behavior rather than device-specific details.
2. **Synthesis:** The synthesis tool parses the HDL source, performs logical optimizations, and maps the design onto primitive cells that suit the target FPGA technology. The output is typically a structural **netlist** (e.g., EDIF or structural Verilog) which details how the design's logic is broken down into LUTs, FFs, and other vendor-specific cells.

3. **Placement and Routing (Implementation):** In *placement*, each logical cell from the synthesized netlist is assigned to a physical location on the FPGA fabric. For instance, LUTs and FFs go into specific *BELs* within the device's CLB sites, and specialized cells such as DSPs and Block RAMs must be placed in their corresponding tile types. Next, *routing* determines how signals are physically wired through the FPGA's configurable interconnect network. Modern tools often interleave these steps (e.g., fluid-placement routing or routing-aware

7 RapidWright API

RapidWright is an open-source Java framework from AMD/Xilinx that provides direct access to the netlist and device databases used by vendor tools. This framework positions itself as an additional workflow column, allowing users to intercept or replace stages of the standard design flow with custom optimization stages (see Figure 23).

- **Design Checkpoints:** RapidWright leverages .dcp files (design checkpoints) generated at various stages of a Vivado flow. By importing a checkpoint, engineers can manipulate the netlist, placement, or routing externally, then re-export a modified checkpoint for further processing in the Vivado workflow column.
- **Key Packages:** RapidWright revolves around three primary data model packages:
 1. **edif** – Represents the logical netlist in an abstracted EDIF-like structure.
 2. **design** – Contains data structures for the physical implementation (Cells, Nets, Sites, BELs, etc.).
 3. **device** – Provides a database of the target FPGA architecture (e.g., Site coordinates, Tile definitions, routing resources).
- **Interfacing with the Netlist and Device:** An engineer can query the netlist to find specific resources (LUTs, FFs, DSPs, etc.) and then map or move them onto device sites. This level of control over backend resources is necessary for research in custom placement, advanced packing techniques, or experimental routing algorithms.

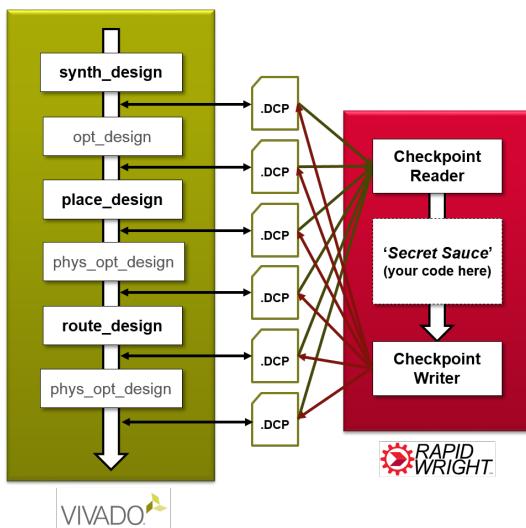


Figure 23: RapidWright workflow integrating into the default Vivado design flow.

By exposing these low-level internals, RapidWright allows fine-grained design transformations that go beyond the standard Vivado IDE's capabilities. Researchers can prototype new EDA strategies without needing to reimplement an entire FPGA backend from scratch, thus accelerating innovation in placement and routing methodologies.

7.1 What is a Netlist?

In its most general form, a netlist is a list of every component in an electronic design paired with a list of nets they connect to. Depending on the abstraction level at hand, these components can be transistors, logic gates, macrocells, or increasingly higher-level modules. Generally, a net denotes any group of two or more interconnected components. In an electronics context, a net can be thought of as a wire connecting multiple pins between multiple components, with each wire having one voltage source and one or more voltage sinks. Thus, one could express the netlist as a hypergraph, nodes representing components, hyperedges representing wires connecting two or more component. More precisely, these hyperedges connect the ports on the components, not the components themselves, with each component exposing multiple ports.

In FPGA context, the components are logical cells (LUTs, CARRY4s, etc.) or hierarchical cells (Verilog module instances) with pins connected together by wires. In Vivado, a Netlist can be synthesized as a Hierarchical or a Flattened netlist. Figure 24 shows an example a Verilog design with modules instantiated in a hierarchy. Figure 25 shows the design synthesized into a **hierarchical netlist** with **hierarchical cells** and **leaf cells**. The synthesizer attempts to construct the module hierarchy as close to the module instantiation hierarchy defined by the user design entry. Figure 26 shows the same design but synthesized into a **flattened netlist** containing **only leaf cells**.

In either synthesized netlist, the **leaf cells**, (deepest level cells), must necessarily consist only of **primitive cells** from the architecture's primitive cell library (LUT6, FDRE, CARRY4, DSP48E1, etc.). The netlist can be compiled and exported as a purely structural low-level Verilog file, or an Electrician Design Interchange Format (EDIF) file, both describing the netlist explicitly as a list of cell instances connected by a list of wires.

```

module top_level(
    input wire clk, // clock
    input wire rst, // reset
    input wire ce, // clock enable
    input wire dina, // data in a
    input wire dibn, // data in b
    input wire dinc, // data in c
    output wire[2:0] dout // data out
);
    module_0 m0(
        clk, rst, ce,
        dina, dibn, dinc, dout[0]
    );
    module_1 m1(
        clk, rst, ce,
        dina, dibn, dinc, dout[2:1]
    );
endmodule

```

```

module module_0(
    input wire clk, // clock
    input wire rst, // reset
    input wire ce, // clock enable
    input wire dina, // data in a
    input wire dibn, // data in b
    input wire dinc, // data in c
    output wire reg dout // data out
);
    wire q_0;
    reg q_1;
    assign q_0 = (dina & !dibn) || dinc;
    always @(posedge clk) begin
        if (rst) begin
            q_1 <= 1'b0;
        end else if (ce) begin
            q_1 <= q_0;
            dout <= q_1;
        end
    end
endmodule

```

```

module module_1(
    input wire clk, // clock
    input wire rst, // reset
    input wire ce, // clock enable
    input wire dina, // data in a
    input wire dibn, // data in b
    input wire dinc, // data in c
    output wire reg dout // data out
);
    module_2 m2(
        clk, rst, ce,
        dina, dibn, dinc, dout[0]
    );
    module_3 m3(
        clk, rst, ce,
        dina, dibn, dinc, dout[1]
    );
endmodule

```

```

module module_2(
    input wire clk, // clock
    input wire rst, // reset
    input wire ce, // clock enable
    input wire dina, // data in a
    input wire dibn, // data in b
    input wire dinc, // data in c
    output wire reg dout // data out
);
    wire q;
    assign q = (dina & dibn) || !dinc;
    always @ (posedge clk) begin
        if (rst) begin
            dout <= 1'b0;
        end else if (ce) begin
            dout <= q;
        end
    end
endmodule

```

```

module module_3(
    input wire clk, // clock
    input wire rst, // reset
    input wire ce, // clock enable
    input wire dina, // data in a
    input wire dibn, // data in b
    input wire dinc, // data in c
    output wire reg dout // data out
);
    wire q;
    assign q = (dina & dibn) || !dinc;
    always @ (posedge clk) begin
        if (rst) begin
            dout <= 1'b0;
        end else if (ce) begin
            dout <= q;
        end
    end
endmodule

```

Figure 24: A simple HDL design with module hierarchy.

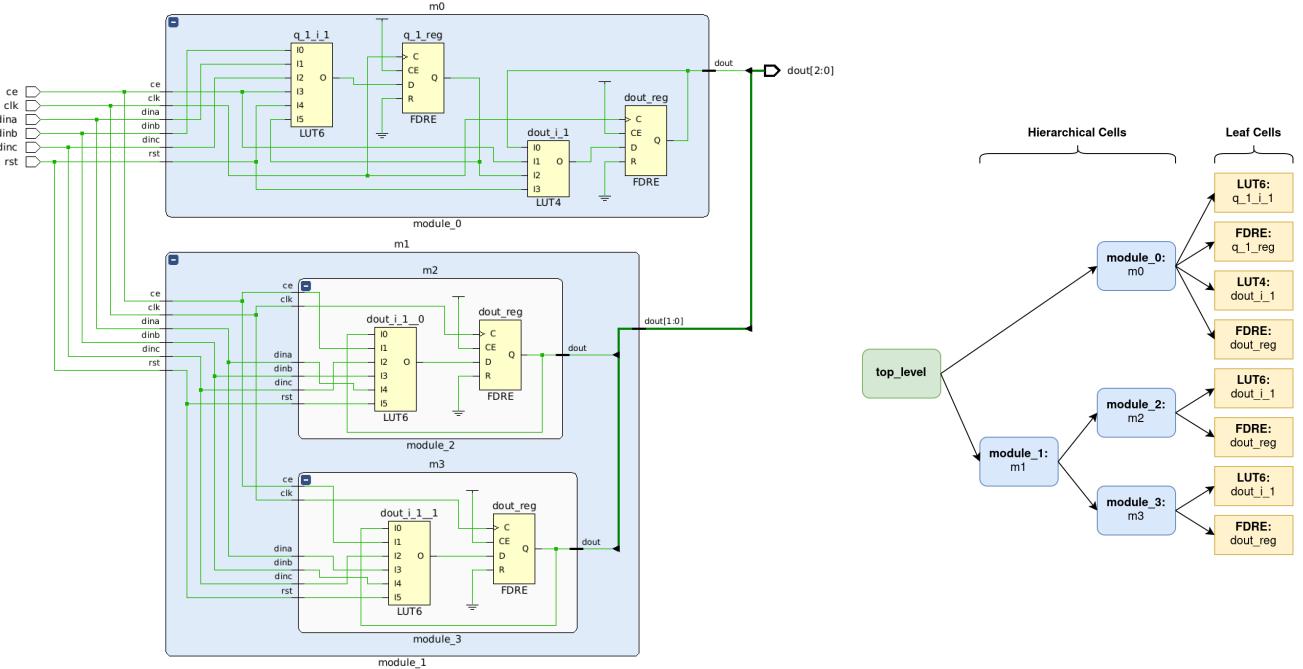


Figure 25: **Left:** A hierarchical netlist consisting of LUTs and FFs. **Right:** The cell hierarchy tree.

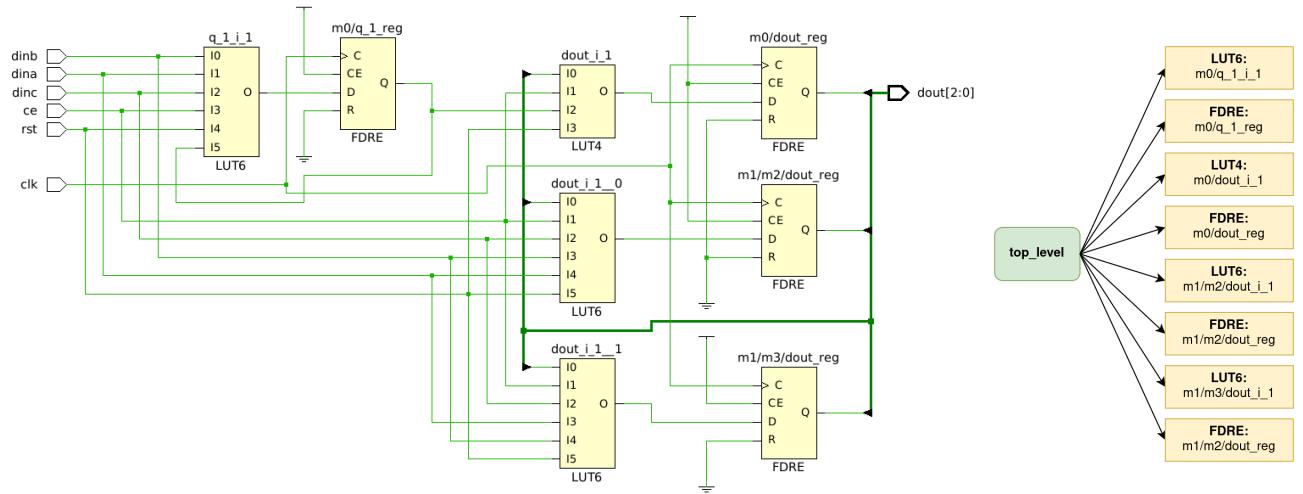


Figure 26: **Left:** A flattened netlist consisting of LUTs and FFs. **Right:** The flattened cell hierarchy tree.

7.2 Netlist Traversal and Manipulation via RapidWright

RapidWright represents the logical netlist objects via the `edif` classes:

- `EDIFNetlist`: The full logical netlist of a Design.
- `EDIFNet`: A logical net within an `EDIFNetlist`.
- `EDIFHierNet`: Combines an `EDIFNet` with a full hierarchical instance name to uniquely identify a net in a netlist.
- `EDIFCell`: A logical cell in an `EDIFNetlist`.
- `EDIFCellInst`: An instance of an `EDIFCell`.
- `EDIFHierCellInst`: An `EDIFCellInst` with its hierarchy, described by all the `EDIFCellInsts` that sit above it within the netlist.
- `EDIFPort`: A port on an `EDIFCell`.
- `EDIFPortInst`: An instance of a port on an `EDIFCellInst`.
- `EDIFHierPortInst`: Combines an `EDIFPortInst` with a full hierarchical instance name to uniquely identify a port instance in a netlist.

Using these classes and their associated methods, we can traverse the logical netlist (`EDIFNetlist`) and analyze or manipulate it as we see fit. A netlist can be easily extracted from a `.dcp` design checkpoint file and traversed like shown in Listing 1. This is performed on the same design shown in figure 26.

Listing 1: Basic netlist extraction and traversal

```

1 Design design = Design.readCheckpoint("synth.dcp")
2 EDIFNetlist netlist = design.getNetlist();
3
4 // Example task:
5 // Extract the set of all unique nets from the design.
6
7 // Initialize a new Set:
8 Set<EDIFNet> netSet = new HashSet<>();
9
10 // Access all leaf cells
11 List<EDIFCellInst> ecis = netlist.getAllLeafCellInstances();
12
13 // Traverse the cell list
14 for (EDIFCellInst eci : ecis) {
15     // Access the ports on this cell
16     Collection<EDIFPortInst> epis = eci.getPortInsts();
17     for (EDIFPortInst epi : epis) {
18         // Access the net on this port
19         EDIFNet net = epi.getNet();
20         netSet.add(net);
21     }
22 }
23
24 // Downstream task:
25 // For each unique net, print out the incident cells.
26
27 // Traverse the set of nets
28 for (EDIFNet net : netSet) {
29     System.out.println("Net: " + net.getName());
30     // Access the ports connected to this net
31     Collection<EDIFPortInst> epis = net.getPortInsts();
32     for (EDIFPortInst epi : epis) {
33         // Access the cell that this port belongs to
34         EDIFCellInst eci = epi.getCellInst();
35         if (eci == null) {
36             // (top_level ports have no associated cell)
37             continue;
38         } else {
39             System.out.println(
40                 "\tCell: " + eci.getName() +
41                 ",\tCellType: " + eci.getCellName()
42             );
43         }
44     }
45 }
```

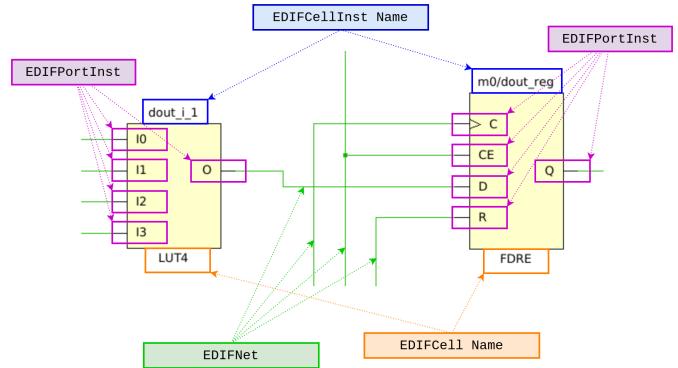


Figure 27: Netlist traversal via the `EDIFCellInst`, `EDIFPortInst`, and `EDIFNet` classes

In this method, we get the list of all `EDIFCellInst` from the `EDIFNetlist` and traverse the list of cells and access their connected nets. Alternatively, we could have also extracted the list of all `EDIFNets` directly from the `design` and traversed the nets to access their connected cells.

Listing 2: Code Printout

```

1 Net: dout[0]
2     Port: I0, Cell: dout_i_1, CellType: LUT4
3     Port: Q, Cell: m0/dout_reg, CellType: FDRE
4 Net: q_1
5     Port: I2, Cell: dout_i_1, CellType: LUT4
6     Port: Q, Cell: m0/q_1_reg, CellType: FDRE
7     Port: I5, Cell: q_1_i_1, CellType: LUT6
8 Net: ce
9     Port: I1, Cell: dout_i_1, CellType: LUT4
10    Port: I1, Cell: dout_i_1_0, CellType: LUT6
11    Port: I1, Cell: dout_i_1_1, CellType: LUT6
12    Port: I3, Cell: q_1_i_1, CellType: LUT6
13 Net: dout[1]
14    Port: I0, Cell: dout_i_1_0, CellType: LUT6
15    Port: Q, Cell: m1/m2/dout_reg, CellType: FDRE
16 Net: dout_i_1_1_n_0
17    Port: O, Cell: dout_i_1_1, CellType: LUT6
18    Port: D, Cell: m1/m3/dout_reg, CellType: FDRE
19 Net: clk
20    Port: C, Cell: m0/dout_reg, CellType: FDRE
21    Port: C, Cell: m0/q_1_reg, CellType: FDRE
22    Port: C, Cell: m1/m2/dout_reg, CellType: FDRE
23    Port: C, Cell: m1/m3/dout_reg, CellType: FDRE
24 Net: dout[2]
25    Port: I0, Cell: dout_i_1_1_1, CellType: LUT6
26    Port: Q, Cell: m1/m3/dout_reg, CellType: FDRE
27 Net: <const>
28    Port: G, Cell: GND, CellType: GND
29    Port: R, Cell: m0/dout_reg, CellType: FDRE
30    Port: R, Cell: m0/q_1_reg, CellType: FDRE
31    Port: R, Cell: m1/m2/dout_reg, CellType: FDRE
32    Port: R, Cell: m1/m3/dout_reg, CellType: FDRE
33 Net: <const1>
34    Port: P, Cell: VCC, CellType: VCC
35    Port: CE, Cell: m0/dout_reg, CellType: FDRE
36    Port: CE, Cell: m0/q_1_reg, CellType: FDRE
37    Port: CE, Cell: m1/m2/dout_reg, CellType: FDRE
38    Port: CE, Cell: m1/m3/dout_reg, CellType: FDRE
39 Net: dout_i_1_0_n_0
40    Port: O, Cell: dout_i_1_0, CellType: LUT6
41    Port: D, Cell: m1/m2/dout_reg, CellType: FDRE
42 Net: rst
43    Port: I3, Cell: dout_i_1, CellType: LUT4
44    Port: I5, Cell: dout_i_1_0, CellType: LUT6
45    Port: I5, Cell: dout_i_1_1, CellType: LUT6
46    Port: I4, Cell: q_1_i_1, CellType: LUT6
47 Net: dinc
48    Port: I2, Cell: dout_i_1_0, CellType: LUT6
49    Port: I2, Cell: dout_i_1_1, CellType: LUT6
50    Port: I2, Cell: q_1_i_1, CellType: LUT6
51 Net: dinb
52    Port: I3, Cell: dout_i_1_0, CellType: LUT6
53    Port: I4, Cell: dout_i_1_1, CellType: LUT6
54    Port: I0, Cell: q_1_i_1, CellType: LUT6
55 Net: dina
56    Port: I4, Cell: dout_i_1_0, CellType: LUT6
57    Port: I3, Cell: dout_i_1_1, CellType: LUT6
58    Port: I1, Cell: q_1_i_1, CellType: LUT6
59 Net: q_1_i_1_n_0
60    Port: D, Cell: m0/q_1_reg, CellType: FDRE
61    Port: O, Cell: q_1_i_1, CellType: LUT6
62 Net: dout_i_1_n_0
63    Port: O, Cell: dout_i_1, CellType: LUT4
64    Port: D, Cell: m0/dout_reg, CellType: FDRE
```

7.3 Placement Flow via RapidWright



Figure 28: The data classes populated at each substage: `PrepackedDesign`, `PackedDesign`, and `PlacedDesign`.

With a basic understanding of FPGA architecture, design placement, and RapidWright, we have all the necessary pieces to implement our SA placer. Here we outline in detail each substage of our implementation: PrePacking, Packing, and Placement. Shown in Figure 29 is an overview of the placement workflow. Figure 28 shows the data structures of RapidWright objects that are populated at each stage: `PrepackedDesign`, which is a group of data structures around `EDIFHierCellInsts`, `PackedDesign`, which is a group of data structures around `SiteInsts`, and finally, `PlacedDesign`, which is simply captured by the final RapidWright `Design` object.

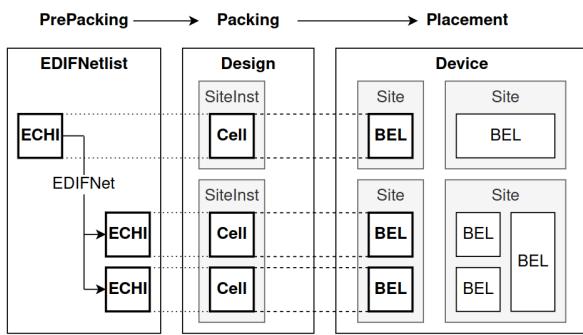


Figure 29: Our placement workflow

8 Prepacking

The first step in our placement flow is **prepacking**. Recall from the 7-Series architecture that there are certain multi-cell structures that must adhere to certain placement constraints to ensure legality, and by design, to minimize wirelength. The job of the prepacker is to traverse the raw EDIF netlist, detect these multi-cell structures, and consolidate these cells into clusters or groups of clusters that naturally reflect these placement constraints.

Recall that CARRY4 chains must necessarily be placed vertically and consecutively across a column of SLICEs in ascending order. Likewise, DSP48E1 cascades must necessarily be placed vertically and consecutively across a column of DSP48E1 Sites in ascending order. A LUT-FF pair may be placed freely, but should be placed in the same lane within the same SLICE to minimize wirelength.

The raw EDIF netlist only tells us the list of nets and the cell ports that they connect to. It does not report the presence of any multi-cell structures (CARRY4 chains, etc.). Thus, we must traverse the netlist to detect these multi-cell structures and store that structure information in a class we will call `PrepackedDesign`.

The code snippet in 3 shows how one can detect and collect these CARRY4 chains in Figure 30 using RapidWright. We first collect the cells in the design that are of type CARRY4, then iteratively traverse their Carry-Out (CO) to Carry-In (CI) nets to find incident CARRY4 cells. Each CARRY4 chain has an anchor cell and a tail cell where the chain terminates. The anchor is found when the CI net connects to Ground (GND), while the tail is found when the CO port is null. We can further detect if there are LUTs or FFs connected to the CARRY4 cell and store that information in a data structure we will call `CarryCellGroup` as defined in figure 28. This will help us in knowing which cells can be packed together into the same `Site` in the subsequent stages.

Similarly, DSP48E1 cascades can be found and collected by traversing the PCOUT ACOUT and BCOUT nets. LUT-FF pairs can be found by inspecting the LUT output (O) net and checking for FF input (DI) ports. We can bucket these LUT-FF pairs by finding the set of unique CE SR net pairs to know which group of LUT-FF pairs can be placed within the same `Site`.

The overall goal of prepacking is to detect the presence of these multi-cellular structures and consolidate that information into our `PrepackedDesign` object in preparation for the following Packing stage.

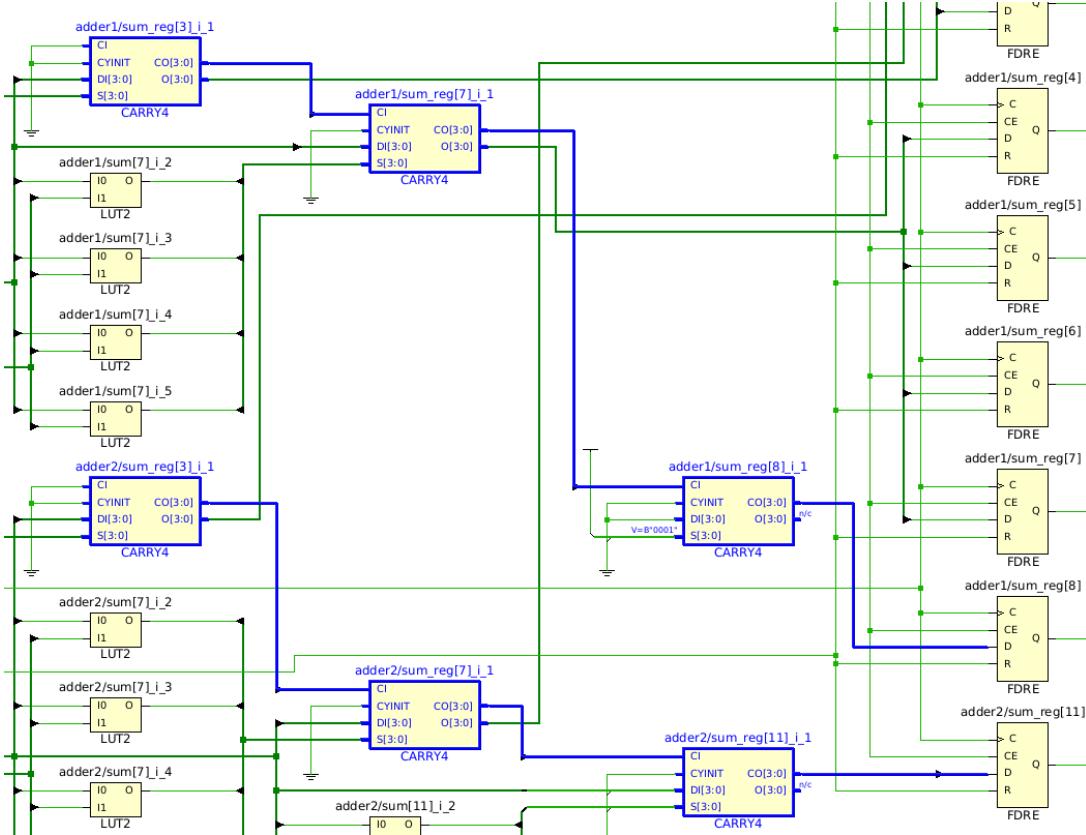


Figure 30: A netlist with two CARRY4 chains, each of size 3

Listing 3: Finding and storing carry chains.

```

1 Design design = Design.readCheckpoint("synth.dcp")
2 EDIFNetlist netlist = design.getNetlist();
3 List<EDIFCellInst> ecis = netlist.getAllLeafCellInstances();
4
5 // Select only the carry cells.
6 List<EDIFCellInst> carryCells = new ArrayList<>();
7 for (EDIFCellInst eci : ecis) {
8     if (eci.getCellName().equals("CARRY4"))
9         carryCells.add(eci);
10 }
11
12 // Find and remove carry chains until the list is empty
13 List<List<EDIFCellInst>> carryChains = new ArrayList<>();
14 while (!carryCells.isEmpty()) {
15     // Set "currentCell" pointer to an arbitrary cell in list
16     EDIFCellInst currentCell = carryCells.get(0);
17
18     // Find this carry chain anchor.
19     // Traverse the Carry-In (CI) to Carry-Out (CO) nets.
20     // Anchor is found when net on the CI Port is Ground.
21     while (true) {
22         System.out.println(currentCell);
23         // Access the CI port on this cell.
24         EDIFPortInst sinkPort = currentCell.getPortInst("CI");
25         // Access the net on this CI port.
26         EDIFNet net = sinkPort.getNet();
27         if (net.isGND()) {
28             break; // Found this chain anchor!
29         }
30         // Get all ports on this net.
31         List<EDIFPortInst> netPorts = net.getPortInsts();
32         for (EDIFPortInst netPort : netPorts) {
33             // Access the port belonging to another carry cell.
34             EDIFCellInst sourceCell = netPort.getCellInst();
35             if (sourceCell.getCellName().equals("CARRY4")) {
36                 // Move the "currentCell" pointer
37                 currentCell = sourceCell;
38                 break;
39             }
40         }
41     }
42
43     // Now currentCell points at this chain's anchor.
44 }
```

```

45     // Now traverse in the opposite direction to find the
46     // chain tail.
47     // Tail is found when the CO Port is null.
48     // Collect the chain cells into an ordered list.
49     List<EDIFCellInst> currentChain = new ArrayList<>();
50     currentChain.add(currentCell);
51     while (true) {
52         EDIFPortInst sourcePort =
53             currentCell.getPortInst("CO[3]");
54         if (sourcePort == null) {
55             break; // Found this chain's tail!
56         }
57         EDIFNet net = sourcePort.getNet();
58         List<EDIFPortInst> netPorts = net.getPortInsts();
59         for (EDIFPortInst netPort : netPorts) {
60             EDIFCellInst sinkCell = netPort.getCellInst();
61             if (netPort.getName().equals("CI") &&
62                 sinkCell.getCellName().equals("CARRY4")) {
63                 currentCell = sinkCell;
64                 // Add the cell to the chain list.
65                 currentChain.add(currentCell);
66                 break;
67             }
68         }
69         // Add currentChain to the list of chains
70         carryChains.add(currentChain);
71         // Remove currentChain from the list of cells
72         carryCells.removeAll(currentChain);
73     } // end while()
74
75     // Print out the carry chains.
76     for (List<EDIFCellInst> chain : chains) {
77         for (int i = 0; i < chain.size(); i++) {
78             EDIFCellInst carry = chain.get(i);
79             if (i == 0) {
80                 writer.write("\nAnchor Cell: " + carry.getName() +
81                             ", CellType: " + carry.getCellName());
82             } else {
83                 writer.write("\n\tCell: " + carry.getName() +
84                             ", CellType: " + carry.getCellName());
85             }
86     }
}
```

Listing 4: Code Printout

```

1 Anchor Cell: adder2/sum_reg[3]_i_1, CellType: CARRY4
2 Cell: adder2/sum_reg[7]_i_1, CellType: CARRY4
3 Cell: adder2/sum_reg[11]_i_1, CellType: CARRY4
4 Anchor Cell: adder1/sum_reg[3]_i_1, CellType: CARRY4
5 Cell: adder1/sum_reg[7]_i_1, CellType: CARRY4
6 Cell: adder1/sum_reg[8]_i_1, CellType: CARRY4
```

9 Packing

Now that we have our `PrepackedDesign` object keeping track of multi-cell structures on the `edif` level, we can start packing them into `SiteInst` objects on the `design` level. Below are some of the most relevant classes from the `design` package for this task.

- **Cell:** A cell corresponds to the leaf cell within the logical netlist `EDIFCellInst` and provides a mapping to a physical location BEL on the device. A cell can be created directly out of an `EDIFCellInst` to inherit all of its `edif` properties on the `design` level.
- **Net:** Represents the physical net to be routed (both inter-site and intra-site). When an `Cell` is created out of an `EDIFCellInst`, the `Nets` are automatically created out of its corresponding `EDIFNets`.
- **SiteInst:** An instance of a `Site` on the `Device`. Carries the mapping information between the `BELs` in a `Site` and the `Cells` assigned to them. Also keeps track of the intra-Site routing information within (`Nets`, `SitePinInsts`, `SitePIPs`, etc.).
- **SitePIP:** A Programmable Interconnect Point (PIP) in a `Site`. Represents the fuses in intra-Site routing `BELs`.
- **SitePinInst:** An instance of a `SitePin` on a `Site`. These objects serve as the interface between intra-Site routing and general inter-Site routing.

A `SiteInst` object can be created and its `BELs` populated with existing `EDIFHierCellInst` objects from the

`EDIFNetlist`. A `SiteInst` is placed on a specific device `Site` upon its creation using the constructor in Listing 5. The packer therefore assigns an initial placement to every `SiteInst` before the simulated-annealing stage randomizes their positions. During this packing phase, we simply map `design SiteInsts` onto `device Sites` in coordinate order as they are generated (e.g., the first `SiteInst` onto `SLICEL_X0Y0`, the second onto `SLICEL_X0Y1`, and so on).

Listing 5 shows how to use the `SiteInst` object and how it ties `Sites`, `Cells`, `BELs`, and intra-Site `Nets` together. The super-specifics of the packing process via RapidWright is too involved for the scope of this paper, but just to give an impression of how the `design` classes are used, Listing 6, 7, 8, and 9 together show how `CarryCellGroups` are packed into `SLICEL` `Sites` in our working code.

Packing the other `Cell` structures into `Sites` follow a similar process. For each cell cluster, we create a `SiteInst` and populate its `BELs` with the corresponding `Cells` in the cluster and assign it a physical `Site` on the device. We further consolidate these `SiteInsts` into clusters or lists of `SiteInsts` for multi-Site structures like `CarryCellGroup` chains and DSP cascades. For each `SiteInst`, we assign it a physical `Site` on the device and route its internal intra-Site nets. The end result is a `PackedDesign` object which is a group of data structures solely around `SiteInsts` like shown in figure 28. The process is analogous to containerization of loose goods in ocean freight forwarding. At this point we no longer need to worry about individual `BELs`, `Cells`, or intra-Site `Nets` as they have now been packaged and taped-off into their respective `SiteInsts`. These `SiteInst` become the new atomic level components in the following Placement stage.

Listing 5: `SiteInst` constructor and methods.

```

1 SiteInst Constructor:
2   SiteInst(String name, Design design, SiteTypeEnum type, Site site)
3 Most relevant SiteInst Methods:
4   createCell(EDIFHierCellInst inst, BEL bel) // Populating the SiteInst BELs with Cells
5   unplace() // Unplacing the SiteInst from its current Site
6   place(Site site) // Placing an unplaced SiteInst onto a Site
7   routeSite() // Attempt to automatically route all intra-Site nets (manual intervention likely required)
8   routeIntraSiteNet(Net net, BELPin src, BELPin snk) // Manually route an intra-Site net
9
10 Example:
11   SiteInst si = new SiteInst("mySiteInst", design, SiteTypeEnum.SLICEL, device.getSite("SLICEL_X0Y1"));
12   si.createCell(someFDRECell, si.getBEL("AFF"));
13   si.routeSite();
14   si.unplace();
15   si.place(device.getSite("SLICEL_X15Y33"));
16   // In Simulated Annealing, SiteInst objects will be unplaced() and placed() many times to converge to an optimal solution.
17   // All Cell-BEL mapping and intra-Site routing is preserved when a SiteInst is moved.

```

Listing 6: Packing one `CarryCellGroup` into one `SLICEL` `SiteInst`.

```

1 // Names of the BELs in the 4 LUT-FF lanes:
2 protected String[] FF_BELS = new String[] { "AFF", "BFF", "CFF", "DFF" };
3 protected String[] LUT6_BELS = new String[] { "A6LUT", "B6LUT", "C6LUT", "D6LUT" };
4 // Pack one CarryCellGroup
5 private void packCarrySite(CarryCellGroup carryCellGroup, SiteInst si) {
6   // Iterate through 4 LUT-FF Lanes
7   for (int i = 0; i < 4; i++) {
8     EDIFHierCellInst ff = carryCellGroup.ffs().get(i);
9     if (ff != null)
10       si.createCell(ff, si.getBEL(FF_BELS[i]));
11     EDIFHierCellInst lut = carryCellGroup.luts().get(i);
12     if (lut != null)
13       si.createCell(lut, si.getBEL(LUT6_BELS[i]));
14     // carry site LUTs MUST be placed on LUT6 BELs.
15     // only LUT6/06 can connect to CARRY4/S0
16   }
17   si.createCell(carryCellGroup.carry(), si.getBEL("CARRY4"));
18   // default intrasite routing
19   si.routeSite();
20   // sometimes the default routeSite() is insufficient, so some manual intervention is required
21   rerouteCarryNets(si);
22   rerouteFFClkSrcCeNets(si);
23 } // end placeCarrySite()

```

Listing 7: Packing CarryCellGroups into SLICEL SiteInsts.

```

1 // Pack all CarryCellGroup chains
2 private List<List<SiteInst>> packCarryChains(List<List<CarryCellGroup>> EDIFCarryChains)
3     throws IOException {
4     List<List<SiteInst>> siteInstChains = new ArrayList<>();
5     writer.write("\n\nPacking carry chains... (" + EDIFCarryChains.size() + ")");
6     for (List<CarryCellGroup> edifChain : EDIFCarryChains) {
7         List<SiteInst> siteInstChain = new ArrayList<>();
8         writer.write("\n\tChain Size: (" + edifChain.size() + "), Chain Anchor: "
9             + edifChain.get(0).carry().getFullHierarchicalInstName());
10        Site anchorSite = selectCarryAnchorSite(edifChain.size());
11        SiteTypeEnum selectedSiteType = anchorSite.getSiteTypeEnum();
12        for (int i = 0; i < edifChain.size(); i++) {
13            Site site = (i == 0) ? anchorSite
14                : device.getSite("SLICE_X" + anchorSite.getInstanceX() + "Y" + (anchorSite.getInstanceY() + i));
15            SiteInst si = new SiteInst(edifChain.get(i).carry().getFullHierarchicalInstName(), design,
16                selectedSiteType,
17                site);
18            packCarrySite(edifChain.get(i), si);
19            if (i == 0) { // additional routing logic for anchor site
20                Net CINNet = si.getNetFromSiteWire("CIN");
21                CINNet.removePin(si.getSitePinInst("CIN"));
22                si.addSitePIP(si.getSitePIP("PRECYINIT", "0"));
23            }
24            occupiedSites.get(selectedSiteType).add(site);
25            availableSites.get(selectedSiteType).remove(site);
26            siteInstChain.add(si);
27        }
28        siteInstChains.add(siteInstChain);
29    } // end for (List<EDIFCellInst> chain : EDIFCarryChains)
30    return siteInstChains;
31 } // end packCarryChains()

```

Listing 8: Manually rerouting intra-Site nets in a SLICEL containing a CARRY4 Cell

```

1 protected String[] FF_BELS = new String[] { "AFF", "BFF", "CFF", "DFF" };
2 private void rerouteCarryNets(SiteInst si) {
3     // activate PIPs for CARRY4/COUT
4     si.addSitePIP(si.getSitePIP("COUTUSED", "0"));
5     // undo default CARRY4/DI nets
6     SitePinInst AX = si.getSitePinInst("AX");
7     if (AX != null)
8         si.unrouteIntraSiteNet(AX.getBELPin(), si.getBELPin("ACYO", "AX"));
9     SitePinInst DX = si.getSitePinInst("DX");
10    if (DX != null)
11        si.unrouteIntraSiteNet(DX.getBELPin(), si.getBELPin("DCYO", "DX"));
12    // activate PIPs for CARRY4/DI pins
13    si.addSitePIP(si.getSitePIP("DCYO", "DX"));
14    si.addSitePIP(si.getSitePIP("CCYO", "CX"));
15    si.addSitePIP(si.getSitePIP("BCYO", "BX"));
16    si.addSitePIP(si.getSitePIP("ACYO", "AX"));
17    // remove stray CARRY4/CO nets
18    if (si.getNetFromSiteWire("CARRY4_CO2") != null)
19        design.removeNet(si.getNetFromSiteWire("CARRY4_CO2"));
20    if (si.getNetFromSiteWire("CARRY4_CO1") != null)
21        design.removeNet(si.getNetFromSiteWire("CARRY4_CO1"));
22    if (si.getNetFromSiteWire("CARRY4_CO0") != null)
23        design.removeNet(si.getNetFromSiteWire("CARRY4_CO0"));
24    // add default XOR PIPs for unused FFs
25    for (String FF : FF_BELS)
26        if (si.getCell(FF) == null)
27            si.addSitePIP(si.getSitePIP(FF.charAt(0) + "OUTMUX", "XOR"));
28 } // end rerouteCarryNets()

```

Listing 9: Manually reroute intra-Site nets in a SLICEL containing FF Cells

```

1 private void rerouteFFClkSrCeNets(SiteInst si) {
2     si.addSitePIP("CLKINV", "CLK");
3     // activate PIPs for SR and CE pins
4     Net SRNet = si.getNetFromSiteWire("SRUSEDMUX_OUT");
5     Net CENet = si.getNetFromSiteWire("CEUSEDMUX_OUT");
6     // if routeSite() default PIPs are incorrect, deactivate them then activate correct PIP
7     if (SRNet != null) {
8         if (SRNet.isGNDNet()) {
9             if (si.getSitePIPStatus(si.getSitePIP("SRUSEDMUX", "IN")) == SitePIPStatus.ON)
10                 for (String FF : FF_BELS)
11                     si.unrouteIntraSiteNet(si.getSitePinInst("SR").getBELPin(), si.getBELPin(FF, "SR"));
12                 si.addSitePIP(si.getSitePIP("SRUSEDMUX", "0"));
13             } else {
14                 if (si.getSitePIPStatus(si.getSitePIP("SRUSEDMUX", "0")) == SitePIPStatus.ON)
15                     for (String FF : FF_BELS)
16                         si.unrouteIntraSiteNet(si.getBELPin("SRUSEDGND", "0"), si.getBELPin(FF, "SR"));
17                     si.addSitePIP(si.getSitePIP("SRUSEDMUX", "IN"));
18             }
19         }
20         if (CENet != null) {
21             if (CENet.isVCCNet()) {
22                 if (si.getSitePIPStatus(si.getSitePIP("CEUSEDMUX", "IN")) == SitePIPStatus.ON)
23                     for (String FF : FF_BELS)
24                         si.unrouteIntraSiteNet(si.getSitePinInst("CE").getBELPin(), si.getBELPin(FF, "CE"));
25                     si.addSitePIP(si.getSitePIP("CEUSEDMUX", "1"));
26             } else {
27                 if (si.getSitePIPStatus(si.getSitePIP("CEUSEDMUX", "1")) == SitePIPStatus.ON)
28                     for (String FF : FF_BELS)
29                         si.unrouteIntraSiteNet(si.getBELPin("CEUSEDGND", "1"), si.getBELPin(FF, "CE"));
30                     si.addSitePIP(si.getSitePIP("CEUSEDMUX", "IN"));
31             }
32         }
33 } // end rerouteFFSrCeNets()

```

10 Placement

So far we have only dealt with technology problems where we consider the problem space and address the constraints of the hardware architecture. For every constraint, we devise a solution to fully address it or a simplification that sacrifices some optimization but reduces the complexity of the problem space.

Now that we have dealt with the bulk of these constraints and squared them away into self-contained `SiteInsts`, we can now more easily apply some general optimization algorithms to finally address our optimization objective, which is to place these `SiteInst` objects onto the device while minimizing wirelength.

10.1 Simulated Annealing

In this paper we implement and present a basic Simulated Annealing (SA) placer. As mentioned in the Introduction, SA is a metaheuristic that approximates a global optimum in a large optimization search space. SA does not guarantee a globally optimal solution, but provides solutions that are often "good enough", especially when finding an approximate global optimum is more important than finding a precise local optimum in a fixed amount of time. SA is typically employed in discrete combinatorial optimization problems like the travelling salesman problem or job-shop scheduling. It is directly inspired by Annealing in metallurgy, which is the physical process of heating a metal above its crystallization temperature, allowing its atoms to migrate through the crystal lattice, and slowly cooling it to allow its atoms to recrystallize into a more desirable structure, such as one that minimizes structural defects.

The state s of the system and the function $E(s)$ to be minimized is analogous to the internal energy of the system. In metallurgic Annealing, the state s represents the position of every atom in the metal object at any given snapshot in time, while $E(s)$ represents the defectiveness in the metal's structure however that may be quantified. In the context of our SA placer, the state s represents the placement of every `SiteInst` on the device at any given iteration, while $E(s)$ represents the total wirelength of `Nets` between all `SiteInsts` of the current placement. We will be measuring wirelength using the manhattan distance between the 2-dimensional coordinates of the `SiteInsts` on the device. We will refer to these distances as half perimeter wirelength (HPWL). The purpose of SA is to bring the system from some initial state to a state with lower energy until some stopping criterion is met, whether that be when a computational budget is exceeded, an energy budget is met, or until the energy's rate of change approaches zero.

At each step or iteration, the algorithm considers some neighboring state s^* . If s^* has a lower energy than s , then the state transition from s to s^* is accepted outright. If s^* has a higher energy, then it can probabilistically be accepted depending on the current global temperature. The higher the temperature, the higher the likelihood of accepting an energy-increasing move. The global temperature starts at some positive amount then gradually decreases to zero according to a cooling schedule. This cooling schedule can be linear, geometric, logarithmic, or some piecewise combination. Its parameters can be fine tuned through empirical experimentation.

Shown in Listing 10 is a simplified pseudocode for the outer-most loop of our placer's SA algorithm. First, we randomly place all `SiteInsts` across the device. Then, we initialize a simple geometric cooling schedule 1:

$$T_{n+1} = \alpha T_n \quad (1)$$

with initial temperature $T_0 = 10000$, cooling rate $\alpha = 0.98$, and $n = \{0, 1, 2, \dots, 300\}$ passes. Hill-climbing moves can be accepted via the probability function in 2:

$$P(s, s^*, T_n) = \exp\left(\frac{E(s) - E(s^*)}{T_n}\right) \quad (2)$$

where $E(s^*)$ is the energy of the proposed state s^* , $E(s)$ is the energy of the current state s , and T_n is the current global temperature at n passes. If $P(s, s^*, T_n) < \text{random}(0, 1)$, then $s \leftarrow s^*$. Recall that this probability is only evaluated for hill-climbing moves, thus the cost difference $E(s) - E(s^*)$ is always negative. In summary, a lower cost difference and a higher temperature contributes to a higher probability of hill-climbing acceptance. All hill-descending moves on the other hand are simply accepted outright.

Then, we enter the outer-most loop which terminates when the number of passes exceeds our computational budget, in this case, 300 passes. In each pass, we `moveAll()` the design, then update the global temperature.

In each `moveAll()`, we iterate through four mutually exclusive groups of placement objects: `DSPSiteInstCascades`, `CARRYSiteInstChains`, `RAMSiteInsts`, and `CLBSiteInsts` (single SLICEs that do not belong to CARRY chains). For each placement object in the pass, we propose a new placement for that object on the device. Each proposed movement represents a proposed state transition of system. If the design has for example, 50 `DSPSiteInstCascades`, 100 `CARRYSiteInstChains`, 75 `RAMSiteInsts`, and 250 `CLBSiteInsts`, and we set a budget of 300 passes, then the algorithm will consider up to $300 * (50 + 100 + 75 + 250) = 142500$ state transitions. If the proposed `Site` is already occupied by a resident or existing `SiteInst`, we evaluate a swap in their placements.

In SA, proposed movements are selected randomly, then accepted or rejected based on legality and cost. Figure 31 shows how a swap proposal between singular `CLBSiteInsts` or between `RAMSiteInsts` is evaluated, while Listing 11 shows the corresponding Java code. On the other hand, the movement of `SiteInst` chains is considerably more complex because there are more placement constraints to consider, especially when swapping the positions between multiple chains. Figure 32 shows a demonstration of how two `DSPSiteInstCascades` or two `CARRYSiteInstChains` can be swapped while Listing 13 shows the corresponding pseudocode.

If the move reduces the HPWL cost, then the movement is accepted outright. If the move increases cost, then it can be accepted by chance if the global temperature is high enough to permit the hill-climb. The temperature decreases with each pass until reaching zero, at which point, the algorithm makes exclusively energy-decreasing moves, effectively reducing into a greedy algorithm. The hope is that by this point, the placement has already crystallized into an approximate global optimum and the greediness of the algorithm will help the system settle deeper into

the energy well faster. We can also switch from random selection to centroid selection when the temperature approaches near zero to speed up the settlement further.

Listing 10: SA pseudocode: outer loop

```

1 public void placeDesign(PackedDesign packedDesign) {
2     randomInitialPlacement(packedDesign);
3     // init_temp=10000, alpha=0.98, max_passes=300
4     initCoolingSchedule(10000, 0.98, 300)
5     int passes = 0;
6     while (passes < 300) {
7         updateTemperature(passes);
8         moveAll(packedDesign);
9         passes++;
10    }
11 }
12
13 private void moveAll(PackedDesign packedDesign) {
14     moveSiteChains(packedDesign.DSPSSiteInstCascades);
15     moveSiteChains(packedDesign.CARRYsiteInstChains);
16     moveSingleSite(packedDesign.RAMSiteInsts);
17     moveSingleSite(packedDesign.CLBSiteInsts);
18 }
```

Listing 11: Single Site Movement

```

1 protected void moveSingleSite(List<SiteInst> sites) {
2     for (SiteInst si : sites) {
3         SiteTypeEnum ste = si.getSiteTypeEnum();
4         List<Site> homeConns = findConnectedSites(si, null);
5         Site homeSite = si.getSite();
6         Site awaySite = proposeSite(si, homeConns, true);
7         SiteInst awaySi = occupiedSites.get(ste).get(awaySite);
8         double oldCost = 0;
9         double newCost = 0;
10        if (awaySi != null) {
11            List<Site> awayConns = findConnectedSites(awaySi,
12                null);
13            oldCost += evaluateSite(homeConns, homeSite);
14            oldCost += evaluateSite(awayConns, awaySite);
15            newCost += evaluateSite(homeConns, awaySite);
16            newCost += evaluateSite(awayConns, homeSite);
17        } else {
18            oldCost += evaluateSite(homeConns, homeSite);
19            newCost += evaluateSite(homeConns, awaySite);
20        }
21        if (evaluateMoveAcceptance(oldCost, newCost)) {
22            if (awaySi != null) {
23                unplaceSiteInst(si);
24                unplaceSiteInst(awaySi);
25                placeSiteInst(si, awaySite);
26                placeSiteInst(awaySi, homeSite);
27            } else {
28                unplaceSiteInst(si);
29                placeSiteInst(si, awaySite);
30            }
31        }
32    } // end randomMoveSingleSite()
```

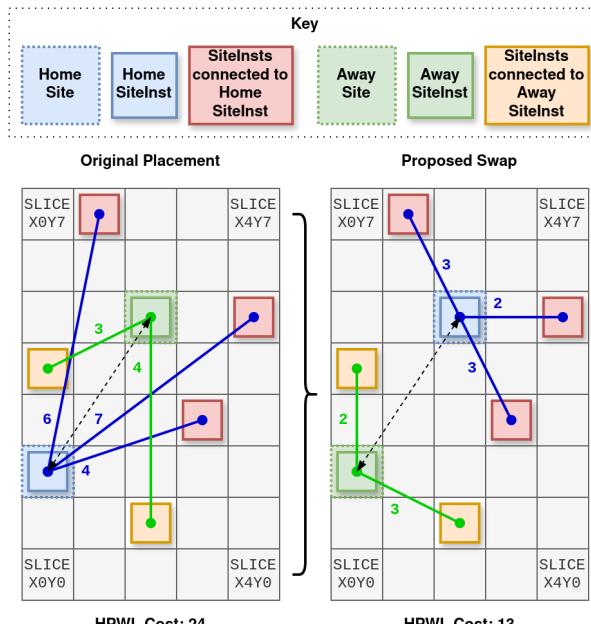


Figure 31: Single Site Swap Proposal

Listing 12: Cooling Schedule and Move Acceptance

```

1 protected List<Double> coolingSchedule;
2 public void initCoolingSchedule(double initialTemp, double
3     alpha, int movesLimit) {
4     double currentTemp = initialTemp;
5     for (int i = 0; i < movesLimit; i++) {
6         this.coolingSchedule.add(currentTemp);
7         currentTemp *= alpha; // geometric cooling
8     }
9 }
10
11 protected boolean evaluateMoveAcceptance(double oldCost,
12     double newCost) {
13     // if the new cost is lower, accept it outright
14     if (newCost < oldCost)
15         return true;
16     // otherwise, evaluate probability to accept higher cost
17     double delta = newCost - oldCost;
18     double acceptanceProbability =
19         Math.exp(-delta / this.currentTemp);
20     return Math.random() < acceptanceProbability;
21 }
```

Listing 13: Chain Swapping Pseudocode

```

1 protected void moveSiteChains(List<List<SiteInst>> chains) {
2     for (List<SiteInst> currentChain : chains) {
3         int chainSize = currentChain.size();
4
5         // Step 1: Identify home window for this chain
6         Site homeAnchor = currentChain.get(0).getSite();
7         List<Site> homeWindow = getSitesInWindow(homeAnchor,
8             chainSize);
9
10        // Step 2: Select a candidate away anchor
11        Site awayAnchor = proposeAnchorSite(currentChain,
12            homeWindow, true);
13
14        // Step 3: Determine away window based on awayAnchor and
15        // chainSize
16        List<Site> awayWindow = getSitesInWindow(awayAnchor,
17            chainSize);
18
19        // Step 4: Find any resident SiteInst chains in the away
20        // window
21        List<List<SiteInst>> residentChainsInAway =
22            collectChainsInWindow(siteType, awayWindow);
23
24        // Step 5: If any resident chains overlap with the away
25        // window, extend the away window to fully accommodate them
26        if (!residentChainsInAway.isEmpty()) {
27            awayWindow = extendWindowToIncludeChains(awayWindow,
28                residentChainsInAway);
29        }
30
31        // Step 6: Map the (possibly extended) away window back
32        // onto the original region so that the tail of that window
33        // coincides with the tail of the current chain
34        List<Site> candidateHomeWindow = mapAwayToHomeWindow(
35            homeAnchor, awayWindow, chainSize);
36
37        // Step 7: While the candidate home window still overlaps
38        // resident chains, shift upward
39        int shifts = 0;
40        while (windowHasOverlap(candidateHomeWindow)) {
41            candidateHomeWindow =
42                shiftWindowUp(candidateHomeWindow);
43            shifts++;
44            if (shifts > (homeWindow.size() - chainSize)) {
45                // Reject this swap attempt.
46                // Move on to the next chain.
47                continue;
48            }
49
50        // Step 8: Compute cost of the original placement in home
51        // window
52        double oldCost = evaluateWindowCost(homeWindow);
53
54        // Step 9: Compute cost of the proposed swap placement
55        double newCost = evaluateWindowCostForSwap(
56            homeWindow, awayWindow, currentChain);
57
58        // Step 10: Decide whether to accept the move based on
59        // oldCost, newCost, and temperature
60        if (evaluateMoveAcceptance(oldCost, newCost, currentTemp))
61        {
62            // Step 11: Perform an element-wise swap of SiteInsts
63            // between homeWindow and awayWindow
64            swapChainsBetweenWindows(homeWindow, awayWindow);
65        }
66    }
67 }
```

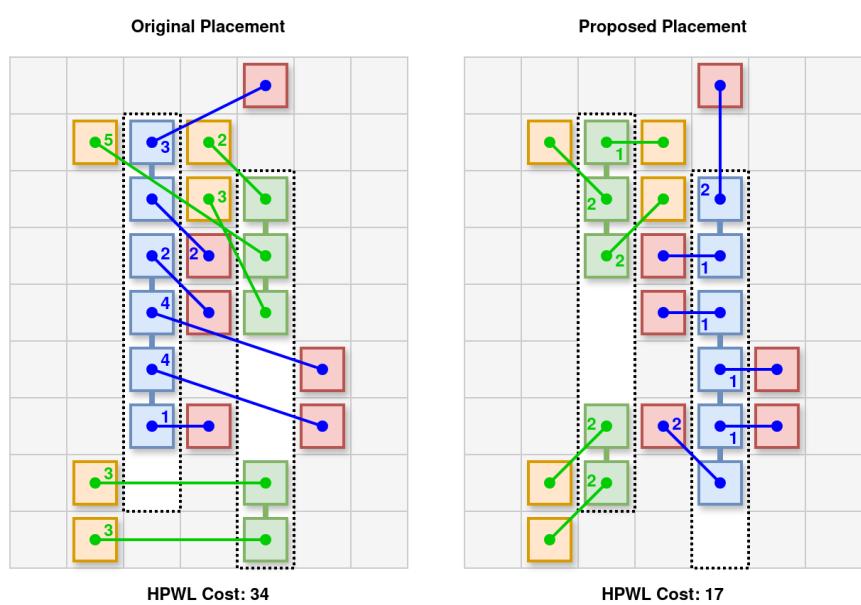
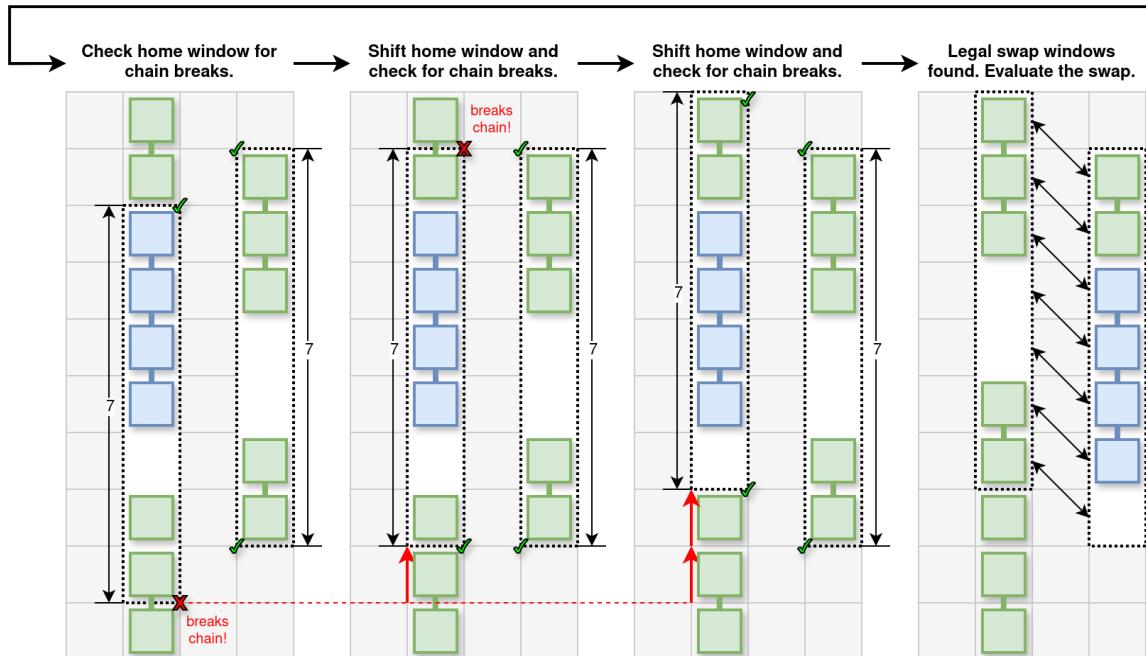
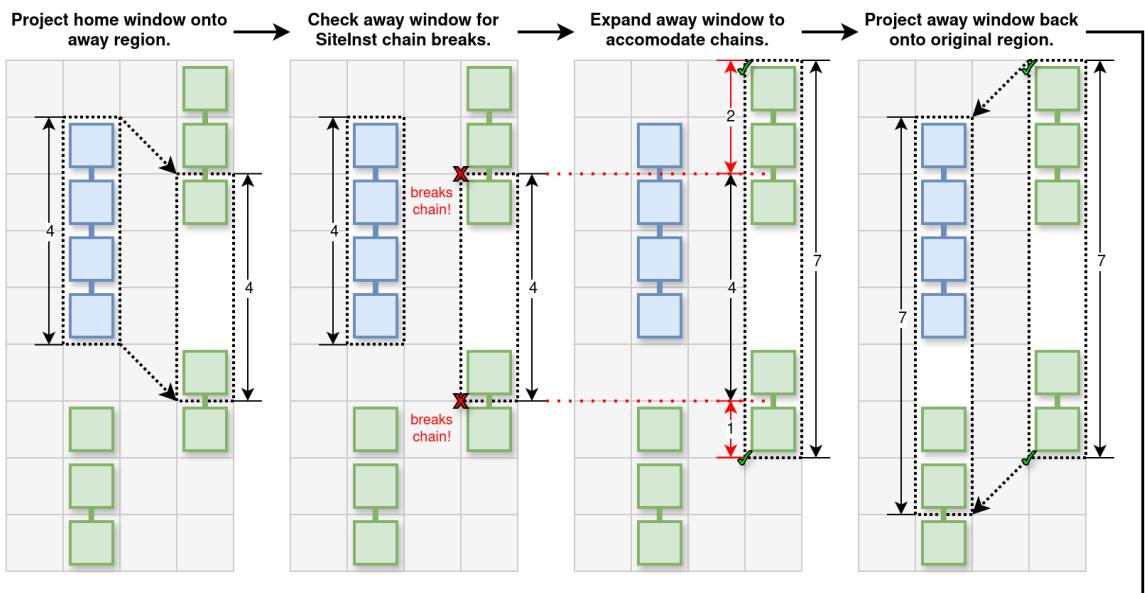


Figure 32: Site Chain Swap Proposal

In more sophisticated placers, the proposed movement can also be selected by finding the centroid between `SiteInsts` that share a net with the current object, or even a hybrid of random and centroid selection. We will implement variations on the vanilla SA algorithm that use centroid selection. Moves that actively attempt to predict a better placement are often referred to as "directed" moves in literature. Placers that take inspiration from physical interactions like gravity or electrostatics are often referred to as "force directed" placers and are akin to physics simulators.

Directed moves as opposed to undirected moves give rise to the concept of legalization. When using undirected moves as in SA, we can simply keep track of a `List` or `HashMap` of occupied `Sites` and available `Sites` and select from that data structure via random selection. However, for directed moves like centroid selection, the centroid between `Sites` may not necessarily fall directly on another discrete `Site` on the `device`. Even when it does, the centroid `Site` may be incompatible with the `SiteInst` or may fail to meet some other placement constraint. In such cases, we must do a neighborhood search around the centroid up to a certain radius to find a `Site` that satisfies all the constraints of the `SiteInst` at hand. We will take some inspiration from ray-marching in graphics and do a diamond-shaped spiral search around the centroid where we check the legality of each `Site` as we step through the spiral and select the first `Site` that meets all of the placement constraints.

In force directed placers, there is a concept of global placement followed by detailed placement. Global placement places object instances freely on an analog plane and is driven purely by directed movement and analytical solution. In global placement stage, only non-hardware related objectives are considered. Object instance placements do not have to adhere to any hardware constraints and can even overlap with one another.

Afterwards, detailed placement "legalizes" the global placement by discretizing the analog positions of the instances onto the discrete grid of resources of the `device`. Overlapping instances are iteratively spread out until every `SiteInst-Cell` is assigned its own unique `Site-BEL`. The spreading process somewhat resembles primitive fluid dynamics simulators or thermal diffusion simulators, where packets of energy are spread through a medium until each voxel or pixel of the medium falls below a maximum threshold of energy. There also exists other strategies to mix packing, placement, and legalization in different orders, such as those studied in [1]. In this context, our placer follows a very straightforward pack-legalize-place strategy with SA.

11 Placement Results

We will need an HDL design that utilizes a good mixture of LUTs, FFs, BRAMs, and DSPs at scale to demonstrate the robustness and performance of our placers. Any DSP subsystem can serve as a good candidate for such a demonstration. Here, we will perform placement on a 2048-order FIR Filter which was conveniently created as coursework for a VLSI course. After passing synthesis, the FIR Filter design calls for the primitive cells shown in the second column of table 33. We will perform placement of this design on a `xc7z020` FPGA, which is a mid-ranged Xilinx device containing the available BELs shown in the first column of table 33. Note that LUT5-LUT6 pairs are counted as a single LUT.

	Available	Usage	Usage (%)
LUT	53200	1693	3
FF	106400	920	1
CARRY4	13300	282	2
RAMB18E1	171	32	23
DSP48E1	220	64	29

Figure 33: Utilization of BELs on a `xc7z020` FPGA for a FIR filter design

Now, we present test results for a basic SA placer as well as test results for four simple variations as listed below. Note that midpoint simply means centroid.

- **PlacerGreedyRandom:** all undirected random moves with greedy acceptance
- **PlacerGreedyMidpoint:** all directed centroid moves with greedy acceptance
- **PlacerAnnealRandom:** all undirected random moves with annealing acceptance (**the basic SA algorithm**)
- **PlacerAnnealMidpoint:** all directed centroid moves with annealing acceptance
- **PlacerAnnealHybrid:** **Initial:** 50% random - 50% centroid moves with annealing acceptance. **At Near-Zero Temperature:** 100% centroid moves with annealing acceptance.

Shown in figure 34 are the HPWL curves for each of the five placers. **PlacerGreedyMidpoint** (in red) appears to perform the worst by a considerable margin with a final HPWL cost of 590K. This is likely due to the lack of randomness or lack of diversity in move proposals combined with the persistent greediness of the movement acceptance that prevents the system from escaping local minima. This conclusion is supported by its HPWL curve (in red) showing a sharp drop in cost that quickly asymptotes just below the 600K mark, suggesting that the system crystallized very quickly into a local minimum and was unable to climb out of it.

Next is **PlacerAnnealMidpoint** with a final cost of 466K. Like the previous placer, **PlacerAnnealMidpoint** lacks the diversity of random moves which contributes to faster crystallization into local minima. However, this placer likely performed better than the previous as there is still some randomness in the move acceptance evaluation which gives way to occasional hill-climbing. Keep in mind that centroid moves do not necessarily decrease

system cost. The centroid move can decrease the HPWL cost of the nets on connected to the current `SiteInst`, but can potentially increase the cost of another group of nets in the event of a swap, which can potentially result in an increase in total system cost.

The next three placers performed better than the previous two placers by decent margin. The vanilla SA placer, `PlacerAnnealRandom`, is represented by the green curve. We can observe that the cost decreases more steadily than the previous two placers, with noticeable spikes in total system cost due to hill-climbing, until settling at the 346K mark as the global temperature cools to zero.

Next is `PlacerGreedyRandom` which is represented by the purple curve. We can observe a much steeper initial drop in system cost than the vanilla SA placer due to greedy movement acceptance. Interestingly, this placer also settled at a lower final system cost of 333K, which is a slight 4% improvement over vanilla SA. We originally expected the greediness to be a detriment to finding the global minimum, but due to the inherent randomness of the algorithm, these broader trends will require many more test trials to reliably observe.

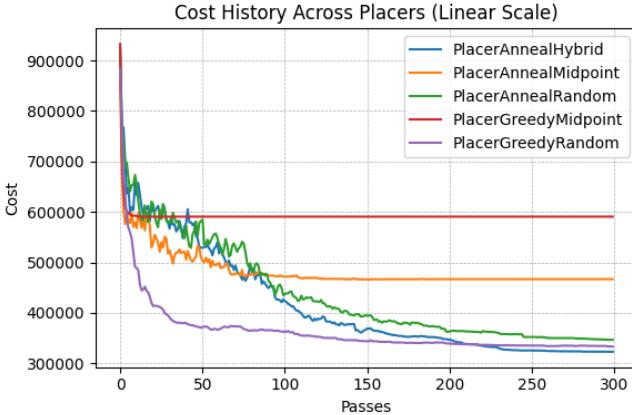


Figure 34: All placers overlaid

The best performing placer was `PlacerAnnealHybrid`, represented by the blue curve. Starting at zero passes, the placer randomly selects between random selection and centroid selection, allowing the placer to make on average more cost-reducing moves, while still maintaining some diversity of moves for hill-climbing. Then, at near-zero temperature, the placer makes exclusively centroid moves with annealing acceptance to speed up the final crystallization into the local minimum. With centroid moves and annealing acceptance, hill-climbing can only be achieved via swaps. We can observe that `AnnealHybrid` settles at a similar rate with the vanilla `AnnealRandom`, but eventually settles at a lower minimum at 322K, slightly outperforming `GreedyRandom`.

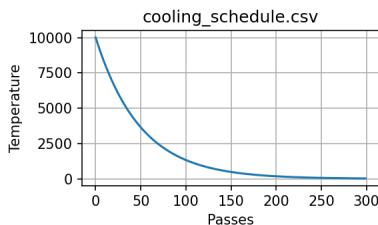


Figure 35: Cooling Schedule: $T_0 = 10000$, $\alpha = 0.99$

Figures 36, 37, 38, 39, and 40 show snapshots of the placement progress on the physical device after 0, 10, 100, and 300 passes for each placer variant.

Please zoom into the document to render the figures in greater detail. All black areas represent purely routing `Sites` that contain no logic resources. All gray or dull-colored boxes represent `Sites` that are unoccupied by `SiteInsts`, while brightly filled boxes represent `Sites` that are occupied by the supported library of `SiteInsts`. Cyan boxes represent occupied `SLICELs`, magenta boxes represent occupied `RAMB18E1s`, and yellow boxes represent occupied `DSP48E1s`.

The tangle of red lines represent the nets between the `Sites`, which are yet to be routed at this stage. Brighter red lines represent nets of higher total HPWL, while duller red lines represent nets of lower total HPWL. Notice there are two brightly red nets in every figure that have a particularly high fanout: one sourced from the near-center, and another sourced from the top right of the device.

The singular white box in the near-center of all of the figures represents an occupied `BUFGCTRL Site`, which is a clock buffer used to route clock nets. Since clock nets tend to have very high fanout, they can use up an unnecessarily large portion of routing resources if they are routed through general routing. Modern FPGAs like the 7-Series have dedicated clock tree routes which can evenly distribute clock signals to all logical `Sites` without burdening the general routing resources. These special clock routes are accessed by routing the raw clock signal from the crystal oscillator IO port to one of these on-chip clock buffers which connects to the on-chip clock tree network. It is always best practice to use them to improve the routability of the design, and simultaneously, to reduce clock skew and improve timing. Our simple FIR Filter design only operates on a single clock domain, so only one clock buffer `Site` is used. An HDL design operating on many clock domains will use more clock buffers or a dedicated clock generator such as a phase locked loop (PLL) or mixed-mode clock manager (MCMM).

The right and left edges of the FPGA device is where most of the general purpose I/O (GPIO) is located. In the top right of the device is where we source our reset signal. Like the clock net, the reset net also has high fanout and is supplied to all synchronous elements. Unlike clock signals, reset signals are not subject to sensitive timing constraints as they are not active on most clock cycles and can tolerate higher skew. Thus, they are simply routed via general routing.

Observe how `PlacerGreedyMidpoint` 37 and `PlacerAnnealMidpoint` 39 both show almost no shrinkage in profile after the first 10 iterations, likely owing to the inability to hill-climb, while the other three placers show incremental shrinkage.

Also note the sparsity of the `DSP48E1` (yellow) and `RAMB18E1` (magenta) `Site` columns compared to the more uniform grid of `SLICELs` (cyan), which poses a computational challenge for the legalization process in directed and analytical placement.

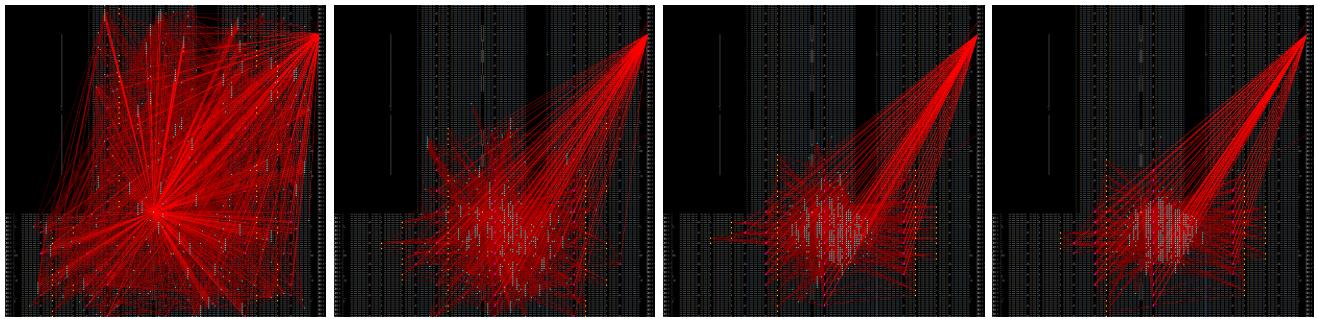


Figure 36: PlacerGreedyRandom

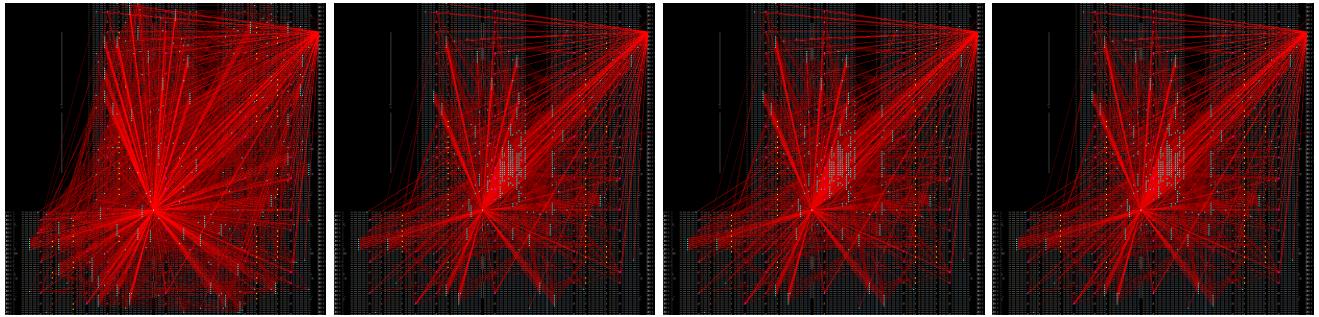


Figure 37: PlacerGreedyMidpoint

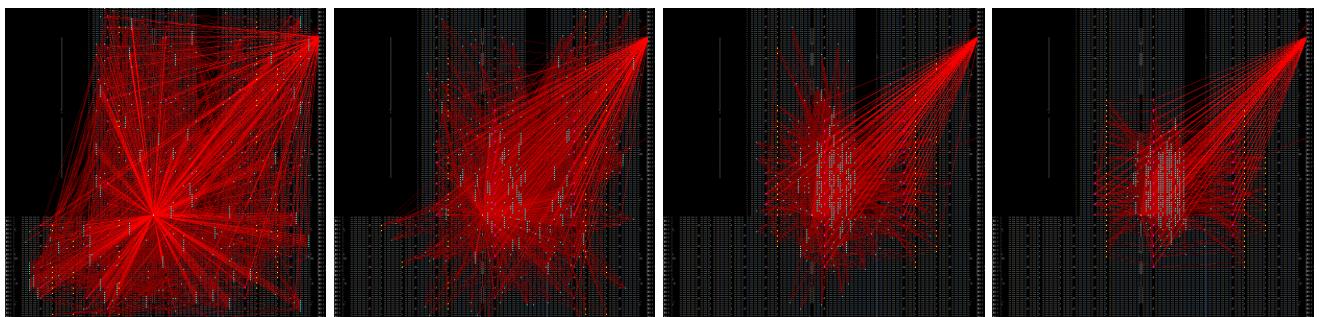


Figure 38: PlacerAnnealRandom

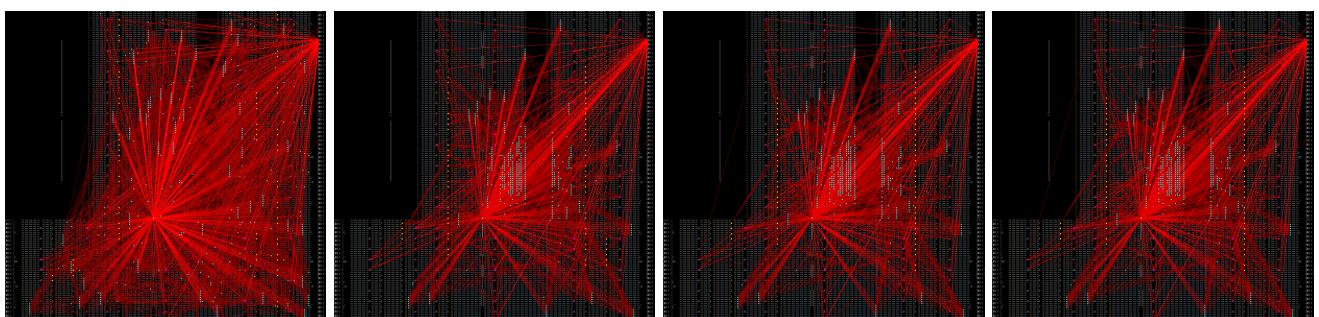


Figure 39: PlacerAnnealMidpoint

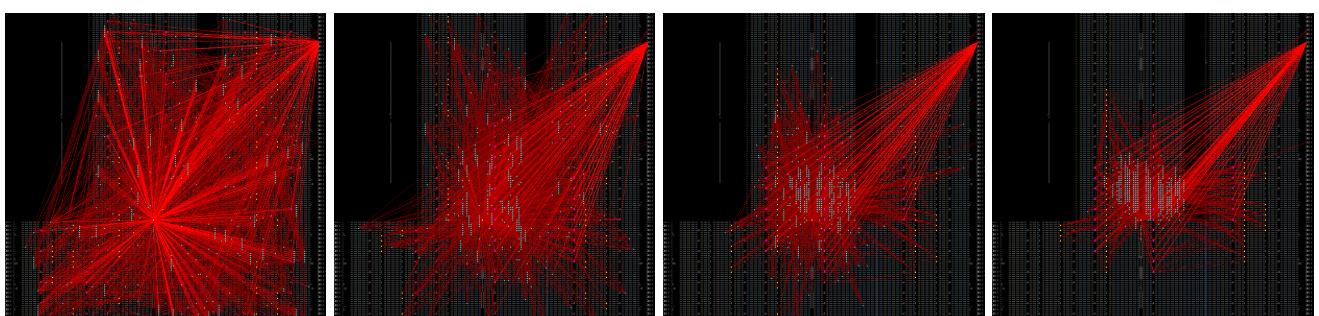


Figure 40: PlacerAnnealHybrid

In addition to variations on the basic SA algorithm, we also test various cooling schedules on the basic SA algorithm. Shown in figure 41 are the cost histories for various cooling schedules, grouped into three different initial

temperatures (5000, 10000, and 15000). For each initial temperature group, we test five different cooling rates (0.79, 0.84, 0.89, 0.94, and 0.99), with the darkness of each curve corresponding with a faster cooling rate.

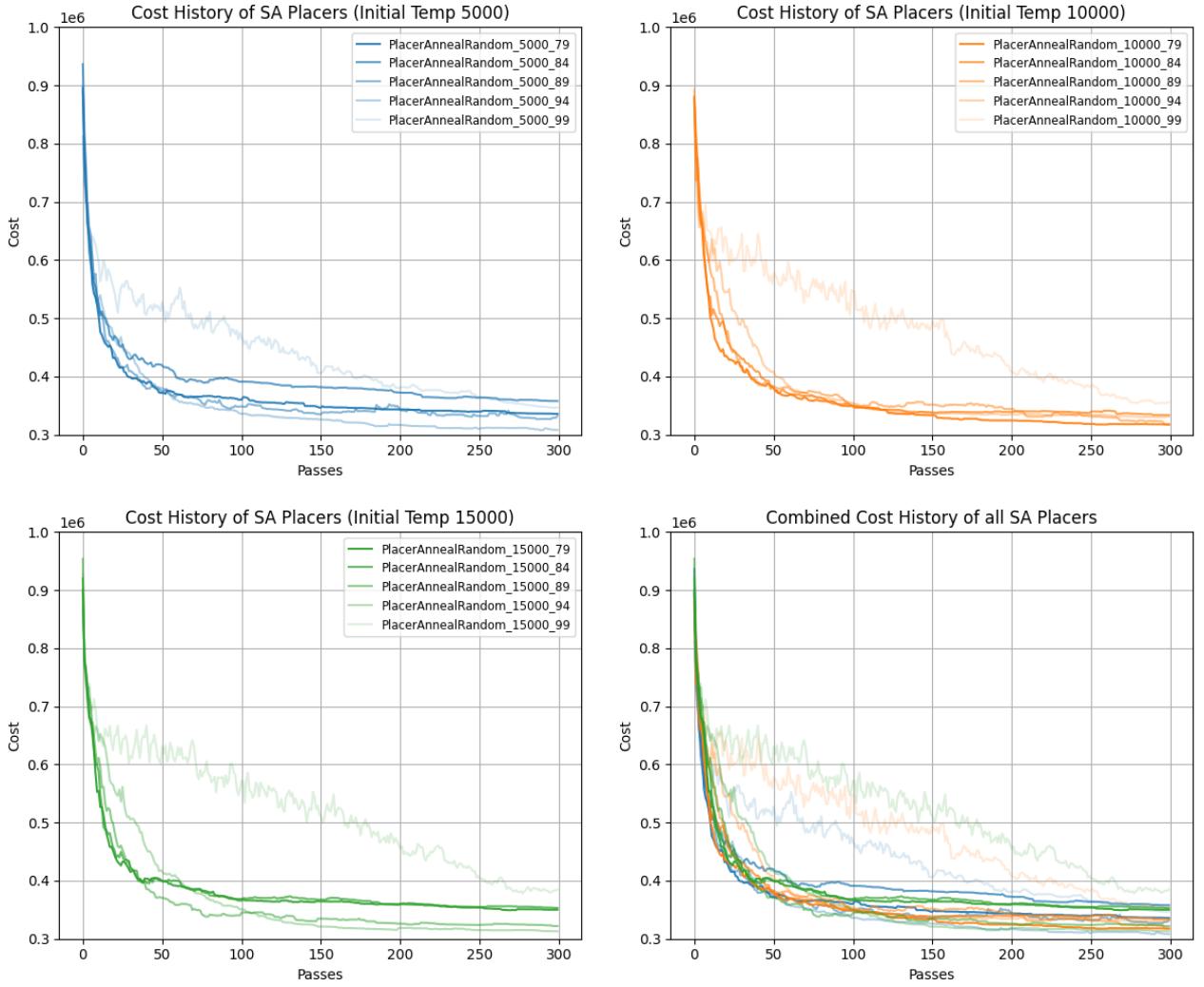


Figure 41: PlacerAnnealRandom with various cooling schedules.

12 Future Work

12.1 Improve Robustness

In this paper we have made many simplifications to the problem space just to make the placer easier to implement. For example, our placer in its current state does not take advantage of SLICEL/SLICEM heterogeneity and simply maps all SLICE SiteInsts onto SLICELs. Recall that the SLICE Sites in Xilinx FPGAs typically come in a 75-25% split between SLICELs and SLICEMs. This means that we have rendered about 25% of the CLB fabric unusable which will inevitably hurt wirelength minimization during placement since the SiteInsts must be spread over a larger area. Enabling SLICEL-SLICEM heterogeneity can lead to greater logic density and consequently less total HPWL, but can make the packing process more complex and may contribute to higher routing congestion.

We can also add packing support for other Xilinx primitives such as the shift-register primitive SRLx, distributed RAM primitives RAMSx, or even LATCH primitives as discussed in 5. Adding support for additional primitives and macros will allow our placer to handle a wider range of

HDL designs and will require deeper consideration of hardware constraints to ensure robustness.

In its current state, the prepacker and packer encounter problems when handling signals larger than 24-bits, especially when involved in DSP functions like addition and multiplication. In such designs, the Vivado synthesizer may synthesize long CARRY4 chains with specific EDIFHierPortInst configurations which are currently not handled by our packer and lead to failures in the subsequent routing stage. Further work is required to resolve these constraints in the packer.

12.2 Analytical Placement

As discussed earlier, simulated annealing (SA) has largely been phased out in state-of-the-art (SOTA) placers due to poor scalability and long runtimes. Our current placer follows a straightforward *prepak–pack–place* flow, with the final placement stage driven by SA. In future work, this stage can be retrofitted to use analytical placement (AP) while reusing the prepacker and packer. Here, SiteInsts remain the atomic placement objects, and analytical solvers can be applied to minimize the total half-

perimeter wirelength (HPWL) of the connecting nets.

The following works provide a foundation for understanding AP, ranging from early formulations to modern SOTA methods:

- Analytical minimization of half-perimeter wirelength — **Kennings and Markov (2000)** [2].
- *Kraftwerk2*—A Fast Force-Directed Quadratic Placement Approach Using an Accurate Net Model — **Spindler et al. (2008)** [3].
- Analytical placement for heterogeneous FPGAs — **Gort et al. (2012)** [4].
- *SimPL*: An algorithm for placing VLSI circuits — **Kim et al. (2013)** [5].
- Multi-Electrostatic FPGA Placement Considering SLICEL-SLICEM Heterogeneity, Clock Feasibility, and Timing Optimization — **Jing et al. (2023)** [6].
- OpenPARF 3.0: Robust Multi-Electrostatics Based FPGA Macro Placement Considering Cascaded Macro Groups and Fence Regions — **Jing et al. (2024)** [7].

As noted in Section 10, replacing SA with AP requires splitting the placement stage into two substages: *global placement*, which determines continuous target positions using analytical optimization, and *detailed placement* (legalization), which snaps objects to valid **sites** while adhering to hardware constraints. Global placement can be performed using any number of off-the-shelf (OTS) analytical solvers, but detailed placement requires custom implementation tailored to the hardware. We repeatedly loop through global placement into detailed placement until HPWL minimization slows to a stop.

Unlike SA, which primarily relies on heuristics and probabilistic moves, AP begins by contextualizing the problem as a collection of mathematical expressions compatible with convex optimization solvers. As briefly introduced in Section 7.1, a circuit netlist can be naturally modeled as a hypergraph

$$G_H = (V_H, E_H) \quad (3)$$

where V_H is the set of vertices and E_H is the set of hyperedges, each of which may connect more than two vertices. In RapidWright context, V_H corresponds to the set of **SiteInsts** placed on the **device**, while E_H corresponds to the set of **Nets** connecting them. More accurately, V_H corresponds to the set of pins (**SitePinInst**, or SPI for short), not the **SiteInsts** themselves. The distinction is usually only important in VLSI context, when the area of a module or macro instance is large enough that the distance between its pins becomes significant [3]. However, in the context of FPGAs where **Sites** are relatively uniform size and whose areas are small compared to the area of the **device**, all **SitePinInsts** belonging to the same **SiteInst** can share the same location coordinates.

This hypergraph can be reduced into a graph using one of several net models:

- **Star model:** classically, introduce a virtual star node and connect it to all pins of the net. However, in electronics, we can simply use the voltage

source node as the star node and avoid adding virtual nodes.

- **Clique model:** connect all pins of the net to each other (quadratic growth in edges).
- **Bound2Bound (B2B) model:** connect only the pins at the geometric boundary (min/max in x or y) to each other, and connect boundary pins to all interior pins. With inverse-distance weights, this model matches HPWL exactly at the point of weight calculation.

Since HPWL is direction-agnostic, directionality of FPGA nets (source/sinks) is typically ignored in reduction. Figure 42 shows an example of hyperedge reduction using the star and clique models while 43 shows the B2B reduction.

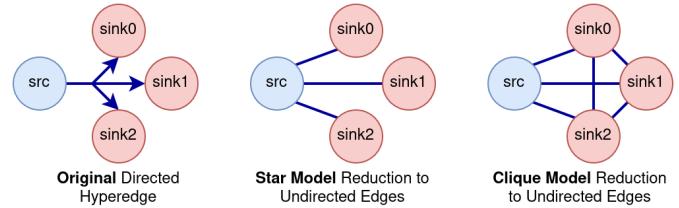


Figure 42: Hyperedge reduction via Star and Clique net models.

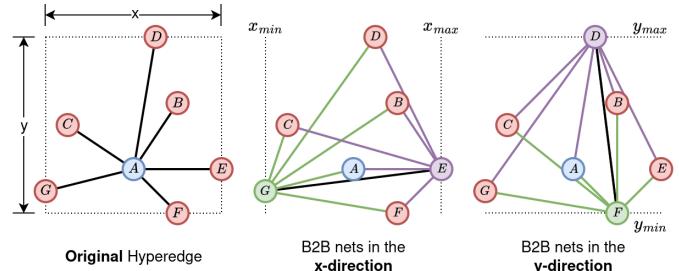


Figure 43: Hyperedge reduction via Bound2Bound net model.

After reduction comes global placement, where AP minimizes a wirelength objective, typically either a linear HPWL formulation (4) or a quadratic approximation (5):

$$\Phi(\mathbf{x}, \mathbf{y}) = \sum_{i,j} w_{ij} (|x_i - x_j| + |y_i - y_j|) \quad (4)$$

$$\Phi(\mathbf{x}, \mathbf{y}) = \sum_{i,j} w_{ij} [(x_i - x_j)^2 + (y_i - y_j)^2] \quad (5)$$

with $i, j \in \{1, 2, \dots, N\}$, where N is the total number of modules in the design. Module placements in the x and y directions are captured by the placement vectors $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)$, where x_n and y_n correspond to the coordinates of the n th module instance. Edges connecting pins i and j have weights w_{ij} and are assigned values depending on the selected net model and objective.

In both formulations, the objective function is separable with x and y . The squared objective is strongly convex and yields a unique minimizer while requiring a single linear solve per iteration. However, it tends to over-emphasize long wires relative to short ones and tends to worsen routing demand compared to linear metrics [2].

The linear wirelength objective more closely reflects how HPWL is measured and better correlates with routability. However, it is non-differentiable, so analytical placers typically handle it by iterated quadratic approximations or by smooth regularizations that enable Newton-type solvers [2]. For ease of implementation, it would be best to first implement the squared objective before exploring the linear objective.

We can use the FPGA placer HeAP (2012) [4], which uses the squared objective along with the B2B net model, as an instructive reference. In HeAP, edges are weighted inversely with the distances between pins and the number of pins on the original hyperedge from which the edge was produced.

$$w_{ij} = \frac{2}{(p-1)|x_i - x_j|}. \quad (6)$$

The squared objective 5 can be recast into matrix-vector form. In the x-direction this can be written as:

$$\Phi(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{Q}_x \mathbf{x} + \mathbf{c}_x^T \mathbf{x} + \text{const}, \quad (7)$$

The connectivity matrix \mathbf{Q}_x encodes 2-pin connections between movable instances while \mathbf{c}_x encodes 2-pin connections between movable instances and fixed instances. Movable instances are for example SLICELs, DSP, and RAM, while fixed instances can be IO pads and clock buffers. We distinguish fixed instances because their placement cannot be optimized and can be thought of as constraints encoded into the objective. For a design with M pins belonging to movable instances and F pins belonging to fixed instances, $\mathbf{Q}_x \in \mathbb{R}^{M \times M}$ and $\mathbf{c}_x \in \mathbb{R}^F$. We can construct \mathbf{Q}_x and \mathbf{c}_x formally by taking the graph Laplacian matrix of our undirected weighted graph (with M-M and M-F connections included), then partitioning the matrix to separate M-M from M-F weights to form \mathbf{Q}_x and \mathbf{c}_x . Alternatively, we can simply follow a shortcut outlined in SimPL [5], where we construct \mathbf{Q}_x by setting the diagonal elements w_{ii} as the sum of net weights of all connections to movable module pin i . The off-diagonal elements w_{ij} are the negative sum of net weights of connections between movable module pins i and j , resulting in a symmetric matrix. The vector \mathbf{c}_x has entry $-c_i$ in the i th row, where c_i is the weight of the connection between fixed pin i and movable pin i , most of which will be zero. The SimPL paper offers a small example of this construction for a toy netlist [5].

Since the objective is strongly convex, we can solve for the global minimum cost by setting the partial derivatives to zero.

$$\frac{\partial}{\partial \mathbf{x}} \Phi(\mathbf{x}) = 0 \iff \mathbf{Q}_x \mathbf{x} = -\mathbf{c}_x \quad (8)$$

Solving this linear system yields the optimal x coordinate placements, and the same applies independently for y . We can use any number of OTS solver frameworks for this step, potentially cvxpy in Python or a Java-native package like EJML. Note that this is only global placement, which is not legal and contains many constraint violations and placement overlaps.

Following global placement is detailed placement (legalization) which is where the real challenge begins. Most placers have a custom method for doing so, tailored to specific optimizations or cost objectives. We can continue

using HeAP as a reference, which performs recursive partitioning to transform the solution provided by the solver into a legal overlap-free placement. The discrete nature of the arrangement of **Sites** on the device and all of the other hardware constraints will make implementation very difficult. Just like in our SA placer, the most challenging parts to implement were the prepacker and packer where we deal with placement constraints, not the placer itself.

12.3 Variations on Packing

Our current placer follows a **Site-centric** approach that resembles that of the old Xilinx ISE - the predecessor to Vivado used in the 2000s. These days, Vivado performs **BEL-centric** placement without necessarily locking **Cells** into **Sites**, allowing for a higher granularity of movement of **Cells**. Enabling BEL-centric movement as opposed to Site-centric movement can improve HPWL minimization, but will add much more complexity to the packing and placement process to ensure robustness with respect to hardware constraints, particularly with intra-Site routing.

There is no constraint that forces us to pack **Cells** into **SiteInsts** before placement or to follow a strict prepacking-packing-placement flow. After implementing a force-directed or analytical placer, we can begin to explore different packing and placement ordering like those studied in Wuxi et al. (2019) [1].

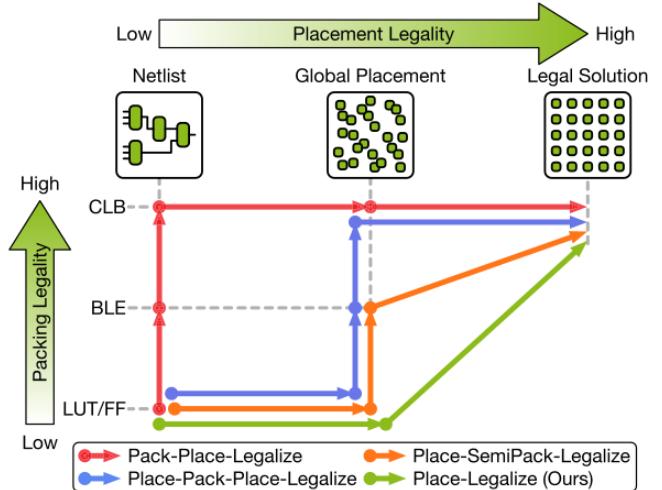


Figure 44: Representative FPGA placement and packing flows. Figure taken from Wuxi et al. (2019), page 1 [1]

12.4 Add Hard Macro Support

As mentioned before, our SA placer is Site-centric. An interesting direction of study would be to expand our placer to support Module-Site-centric mixed placement, where Modules are hard macros comprising multiple Sites. This of course would also add complexity to the prepacker and packer, as now we are dealing with non-uniform mixed-size placement objects. The RapidWright repository has two implementations of purely Module-centric placers, called **BlockPlacer** and **BlockPlacer2**, though it currently lacks a purely Site-centric placer or a mixed Module-Site-centric placer. The corresponding paper by Lavin et. al. (2010) [8] can serve as a useful reference for this direction.

13 Conclusion

This work has demonstrated a complete placement flow for heterogeneous FPGAs built upon the RapidWright framework, culminating in a functional simulated annealing placer. By systematically prepacking multi-cell structures, packing them into SiteInst containers, and applying an SA-driven optimization, we produce placements that respect the discrete architectural constraints of Xilinx 7-Series devices while reducing wirelength. Although SA is computationally expensive compared to modern analytical placers, its simplicity, flexibility, and robustness make it an excellent foundation for research and education. The infrastructure we have developed can be extended in several directions: integrating timing-driven cost functions, incorporating legalization-aware directed moves, and experimenting with analytical or hybrid placement algorithms. Future work will focus on replacing or augmenting the SA stage with an analytical solver to combine the quality of wirelength minimization with the scalability and runtime performance needed for larger industrial designs.

References

- [1] W. Li and D. Z. Pan, “A new paradigm for fpga placement without explicit packing,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 38, no. 11, pp. 2113–2126, 2019.
- [2] A. A. Kennings and I. L. Markov, “Analytical minimization of half-perimeter wirelength,” in *Proceedings of the 2000 Asia and South Pacific Design Automation Conference, ASP-DAC ’00*, (New York, NY, USA), p. 179–184, Association for Computing Machinery, 2000.
- [3] P. Spindler, U. Schlichtmann, and F. M. Johannes, “Kraftwerk2—a fast force-directed quadratic placement approach using an accurate net model,” *Trans. Comp.-Aided Des. Integ. Cir. Sys.*, vol. 27, p. 1398–1411, aug 2008.
- [4] M. Gort and J. H. Anderson, “Analytical placement for heterogeneous fpgas,” in *22nd International Conference on Field Programmable Logic and Applications (FPL)*, pp. 143–150, 2012.
- [5] M.-C. Kim, D.-J. Lee, and I. L. Markov, “Simpl: an algorithm for placing vlsi circuits,” *Commun. ACM*, vol. 56, p. 105–113, June 2013.
- [6] J. Mai, J. Wang, Z. Di, and Y. Lin, “Multi-electrostatic fpga placement considering slicel-slicem heterogeneity, clock feasibility, and timing optimization,” 2023.
- [7] J. Mai, J. Wang, Y. Chen, Z. Guo, X. Jiang, Y. Liang, and Y. Lin, “Openparf 3.0: Robust multi-electrostatics based fpga macro placement considering cascaded macros groups and fence regions,” in *2024 2nd International Symposium of Electronics Design Automation (ISEDA)*, pp. 374–379, 2024.
- [8] C. Lavin, M. Padilla, S. Ghosh, B. Nelson, B. Hutchings, and M. Wirthlin, “Using hard macros to reduce fpga compilation time,” in *2010 International Conference on Field Programmable Logic and Applications*, pp. 438–441, 2010.
- [9] I. L. Markov, J. Hu, and M.-C. Kim, “Progress and challenges in vlsi placement research,” *Proceedings of the IEEE*, vol. 103, no. 11, pp. 1985–2003, 2015.
- [10] A. Mishra, N. Rao, G. Gore, and X. Tang, “Architectural exploration of heterogeneous fpgas for performance enhancement of ml benchmarks,” in *2023 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, pp. 232–235, 2023.
- [11] U. Farooq, H. Parvez, H. Mehrez, and Z. Marrakchi, “Exploration of heterogeneous fpga architectures,” *Int. J. Reconfig. Comput.*, vol. 2011, jan 2011.
- [12] C. Lavin and A. Kaviani, “Rapidwright: Enabling custom crafted implementations for fpgas,” in *2018 IEEE 26th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, pp. 133–140, 2018.
- [13] Y. Zhou, P. Maidee, C. Lavin, A. Kaviani, and D. Stroobandt, “Rwroute: An open-source timing-driven router for commercial fpgas,” *ACM Trans. Reconfigurable Technol. Syst.*, vol. 15, nov 2021.
- [14] C. Lavin and E. Hung, “Invited paper: Rapidwright: Unleashing the full power of fpga technology with domain-specific tooling,” in *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, pp. 1–7, 2023.
- [15] L. Guo, P. Maidee, Y. Zhou, C. Lavin, J. Wang, Y. Chi, W. Qiao, A. Kaviani, Z. Zhang, and J. Cong, “Rapidstream: Parallel physical implementation of fpga hls designs,” in *Proceedings of the 2022 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, FPGA ’22*, (New York, NY, USA), p. 1–12, Association for Computing Machinery, 2022.
- [16] N. Zhang, X. Chen, and N. Kapre, “Rapidlayout: Fast hard block placement of fpga-optimized systolic arrays using evolutionary algorithms,” in *2020 30th International Conference on Field-Programmable Logic and Applications (FPL)*, pp. 145–152, 2020.
- [17] S. Areibi, G. Grewal, D. Banerji, and P. Du, “Hierarchical fpga placement,” *Canadian Journal of Electrical and Computer Engineering*, vol. 32, no. 1, pp. 53–64, 2007.
- [18] J. Chen, W. Zhu, J. Yu, L. He, and Y.-W. Chang, “Analytical placement with 3d poisson’s equation and admm based optimization for large-scale 2.5d heterogeneous fpgas,” in *2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pp. 1–8, 2019.
- [19] S. Dhar, L. Singhal, M. Iyer, and D. Pan, “Fpga accelerated fpga placement,” in *2019 29th International Conference on Field Programmable Logic and Applications (FPL)*, pp. 404–410, 2019.
- [20] D. Shah, E. Hung, C. Wolf, S. Bazanski, D. Gisselquist, and M. Milanovic, “Yosys+nextpnr: An open source framework from verilog to bitstream for commercial fpgas,” in *2019 IEEE 27th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, pp. 1–4, 2019.
- [21] T. Liang, G. Chen, J. Zhao, S. Sinha, and W. Zhang, “Amf-placer: High-performance analytical mixed-size placer for fpga,” in *2021 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pp. 1–6, 2021.
- [22] R. S. Rajarathnam, M. B. Alawieh, Z. Jiang, M. Iyer, and D. Z. Pan, “Dreamplacefpga: An open-source analytical placer for large scale heterogeneous fpgas using deep-learning toolkit,” in *2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 300–306, 2022.
- [23] H. Long, Y. Bai, Y. Li, J. Wang, and J. Lai, “Optimizing wirelength and delay of fpga tile through floorplanning based on simulated annealing algorithm,” in *2023 IEEE 15th International Conference on ASIC (ASICON)*, pp. 1–4, 2023.

- [24] T. Liang, G. Chen, J. Zhao, S. Sinha, and W. Zhang, “Amf-placer 2.0: Open source timing-driven analytical mixed-size placer for large-scale heterogeneous fpga,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, pp. 1–1, 2024.
- [25] E. Hung, F. Eslami, and S. J. Wilton, “Escaping the academic sandbox: Realizing vpr circuits on xilinx devices,” in *2013 IEEE 21st Annual International Symposium on Field-Programmable Custom Computing Machines*, pp. 45–52, 2013.
- [26] B. Ray, A. R. Tripathy, P. Samal, M. Das, and P. Mallik, “Half-perimeter wirelength model for vlsi analytical placement,” in *2014 International Conference on Information Technology*, pp. 287–292, 2014.
- [27] B. N. B. Ray, S. K. Mohanty, D. Sethy, and R. B. Ray, “Hpwl formulation for analytical placement using gaussian error function,” in *2017 International Conference on Information Technology (ICIT)*, pp. 56–61, 2017.
- [28] Y. Lin, S. Dhar, W. Li, H. Ren, B. Khailany, and D. Z. Pan, “Dreampiace: Deep learning toolkit-enabled gpu acceleration for modern vlsi placement,” in *2019 56th ACM/IEEE Design Automation Conference (DAC)*, pp. 1–6, 2019.