International Conference
on Supercomputing 2021
June 14 - 18, 2021. Worldwide online event

POLITECNICO DI MILANO

Bambu: High-Level Synthesis for Parallel Programming

Compiler Based Optimizations, Tuning and Customization of Generated Accelerators

Michele Fiorito

Politecnico di Milano
Dipartimento di Elettronica, Informazione e Bioingegneria
michele.fiorito@polimi.it

Outline

- Reset, registering, FSM encoding
- Tuning accelerators by means of compiler optimizations
- Bambu Optimizations
- System of Difference Constraints
- Math support

Reset types and values

- Internal status of accelerators can be reset
  - Accelerators exposes a reset signal
- Register reset type:
  - no (default)
  - async
  - sync
- Reset level:
  - low (default)
  - high
- Example:

```
--reset-type=sync --reset-level=high
```

Registered inputs

- A dedicated port is created for scalar parameters of each module function
- Generated modules expect stable inputs
  - If inputs are not stable, they can be registered
- Registered inputs:
  - **auto** – (default) inputs are registered only for shared functions
  - **top** – inputs are registered for top interface and shared functions
  - **yes**
  - **no**

```
--registered-inputs=<value>
```

FSM-encoding

- Different types of encoding can be used in Finite State Machine
  - **auto** – (default) depends on target
  - **one-hot**
  - **binary**
- Default: best encoding for logic synthesis tool
  - Vivado: one-hot
  - Other tools: binary

```
--fsm-encoding=<value>
```

Improve Area/Performance of generated accelerators

- Performance and/or area of the generated accelerators can be improved by tuning the design flow
  - GCC/CLANG optimizations
  - Bambu IR optimizations
  - Bambu HLS algorithms
- Best design flow for every accelerator does not exist
  - Trade off between area and performance
  - Effects of the single optimizations can be different on the single accelerators
- Default:
  - Balanced area/performance trade off

GCC/CLANG Optimizations

- Only GCC/CLANG target independent optimizations are considered
- -O3 is not necessarily the best choice
  - Can improve **performances**
  - Can increment **area**
- User can tune this part of the flow:
  - Selecting optimization level:

```
-O0 or -O1 or -O2 or -O3 or -Os
```

  - Enabling/disabling single optimization:

```
-f<optimization> -fno-<optimization>
```

  - Tuning parameters: --param

```
--param <name>=<value>
```

Effect of GCC Optimizations

- Results refer to other Bambu options set to default value

Opts	Cycles	Luts
O0	15764	11675
O1	7892	11052
O2	4679	10276
O3	3854	15679
O3 vectorize	3816	38553
O3 all inline	1327	13550

Bambu IR Analysis

- Collect information used by IR optimizations and High Level Synthesis
- Data flow analysis
  - Scalar: based on SSA
  - Aggregates: exploit GCC+Bambu alias analysis
- Graphs Computation
  - Call Graph, CFG, DFG, …
- Loops identification
- Bit Value Analysis
  - Compute for each SSA which bit are used and which bit are fixed
- Range Analysis

Bambu IR Optimizations

- Applied before HLS to the IR produced by GCC
- Two type of optimizations
  - Single instruction optimizations
  - Multiple instruction optimizations
  - Restructuring of Control Flow Graph
  - Fixing IR
- Sequences of optimizations can be applied multiple times
  - Fixed point iteration optimization flow

Single Instruction Optimization

- IR lowering – make single instructions more suitable to be implemented on FPGA
  - Expansion of multiplication by constant
  - Expansion of division by constant
  - Etc.
- Bit Value Optimization
  - Shrink operations to the only significant bits

Multiple Instruction Optimization

- Common Subexpression Elimination
- Dead Code Elimination
- Extract pattern (e.g., three input sum)
- LUT transformations
  - Merging multiple Boolean operations into a single LUT-based operation
- Conditional Expression Restructuring
- Commutative Expression Restructuring

Restructuring of Control Flow Graph

- Speculation
- Code motion
- Merging of conditional branch
  - Creation of multiple target branch
- Basic Block Manipulation
  - Remove (empty, dead, …)
  - Split
  - Merge

Fixing IR

- Struct assignment
  - Replaced with memcpy call
- Floating point operations
  - Replaced with function calls
- Integer divisions
  - Replaced with function calls

System of Difference Constraints

- Global scheduling based on ILP formulation
- Results are exploited to perform
  - Speculation
  - Code Motion

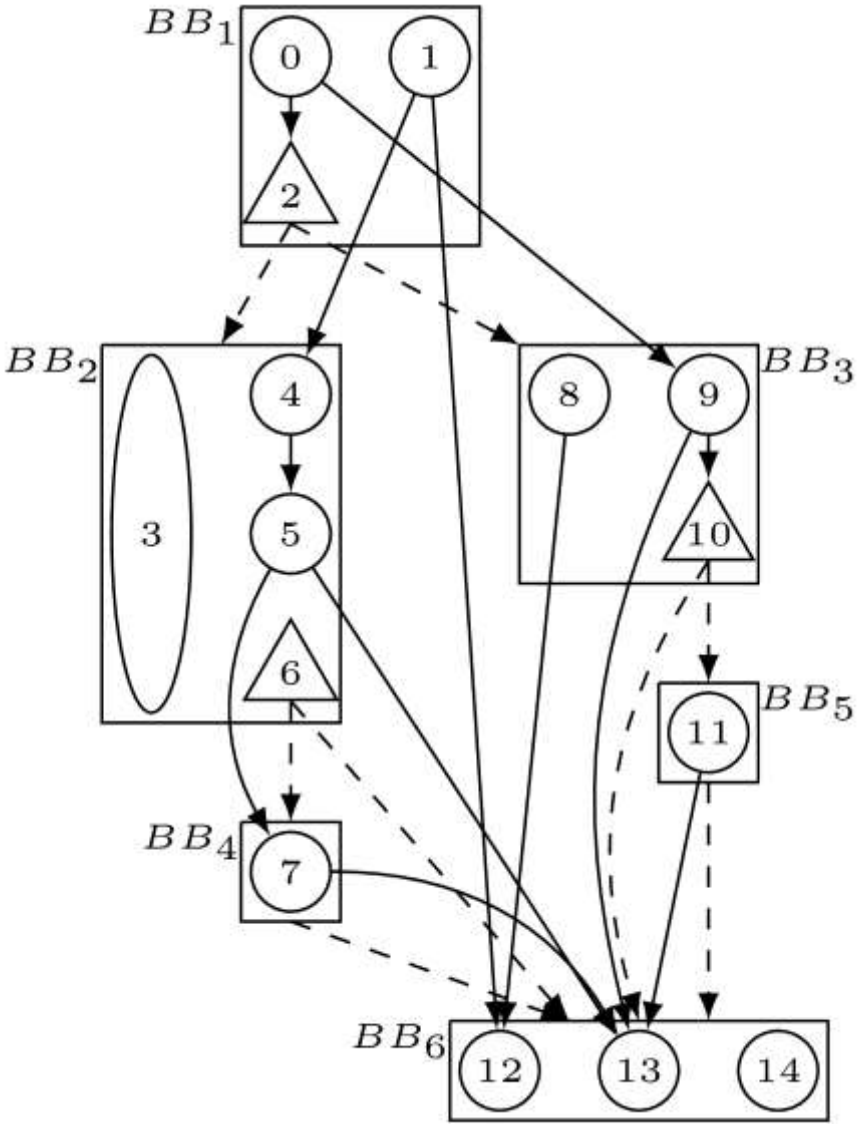
+ Improve performances of accelerators

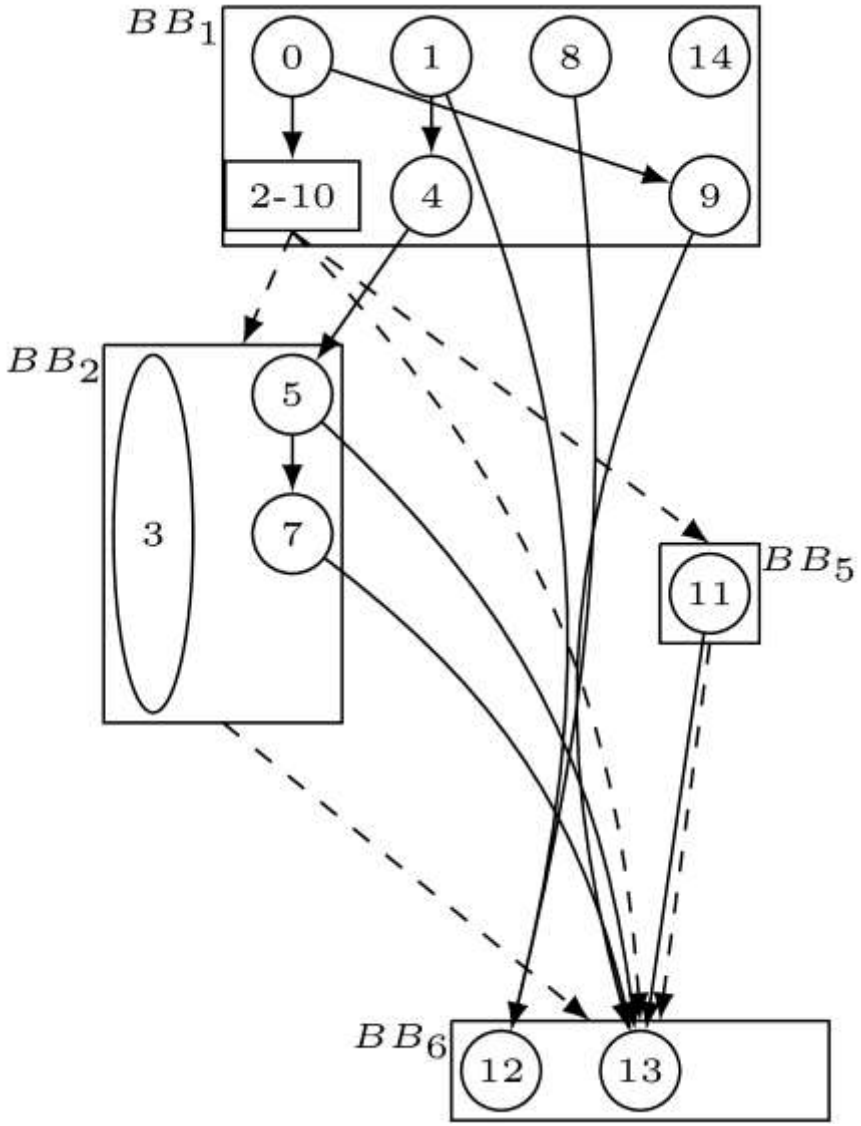
- Potentially increment area of accelerators

- Increase High Level Synthesis time

```
--speculative-sdc-scheduling
```

IR optimizations: Example





Experimental setup

- Predefined design flows

```
--experimental-setup=<setup>
```

BAMBU-AREA: optimized for area

BAMBU-PERFORMANCE: optimized for performances

BAMBU-BALANCED: optimized for trade-off area/performance

BAMBU-AREA-MP, BAMBU-PERFORMANCE-MP, BAMBU-BALANCED-MP: enable support to true dual port memories

Default: BAMBU-BALANCED-MP

Constraints

- Bambu assumes infinite resources during High Level Synthesis
  - Produced solutions may not fit in the target device
- Area of generated solutions can be indirectly controlled by means of **constraints**
- User can constraint the number of available functional units in each function
  - E.g.: fix the number of available multiplier in each function
- Constraints are set by means of *XML file*

Example of constraints file


```
<?xml version="1.0"?>
  <constraints>
    <HLS_constraints>
      <tech_constraints fu_name="mult_expr_FU"
                        fu_library="STD_FU" n="8"/>
    </HLS_constraints>
  </constraints>
```

Translation flow

- C→HDL without optimizations
  - GCC/CLANG optimizations are (mostly) disabled
  - Bambu IR optimizations are (mostly) disabled

```
-O0 --cfg-max-transformations=0 --no-chaining
```

- Can be exploited only when bambu is compiled with development support
- Useful for debugging

Integer Division Algorithms

- You can control how to implement integer divisions:

```
--hls-div=<implementation>
```

- Available implementations:
  - `none`: HDL based pipeline restoring division
  - **`nr1`** (default): C-based non restoring division with unrolling factor equal to 1
  - `nr2`: C-based non restoring division with unrolling factor equal to 2
  - `NR`: C-based Newton-Raphson division
  - `as`: C-based align divisor shift dividentd method

Floating point support

- Possible ways of implementing floating point ops:
  - Softfloat (default): customized faithfully rounded (nearest even) version of soft based implementation

```
--soft-float
```

  - Softfloat-subnormal: soft based implementation with support to subnormal

```
--softfloat-subnormal
```

  - Softfloat GCC: GCC soft based implementation

```
--soft-fp
```

  - Flopoco generated modules

```
--flopoco
```

Libm versions

- Bambu exploits High Level Synthesis to generate accelerators implementing libm functions
- Two different versions of libm are available
  1. Faithfully rounding (default)
  2. Classical libm built integrating existing libm source code from glibc, newlib, uclibc and musl libraries.
     - Worse performances and area

Hands-on time

Switch to Colab Notebook to test some of bambu optimizations

First example – ADPCM

Benchmark	CYCLES	HLS_execution_time
GCC49:adpcm_O0	33429	23,05
GCC49:adpcm_O1	24547	18,72
GCC49:adpcm_O2	24043	43,26
GCC49:adpcm_O3	10429	76,45
GCC49:adpcm_O3_inline	7503	99,58
GCC49:adpcm_O3_vectorize	6995	49,31
GCC49:adpcm_Os	24847	25,21

Second example – ADPCM

Benchmark	CYCLES	HLS_execution_time
GCC49:adpcm_O0_sdc	33479	64,38
GCC49:adpcm_O1_sdc	24297	57,09
GCC49:adpcm_O2_sdc	22863	83,53
GCC49:adpcm_O3_sdc	9149	175,93
GCC49:adpcm_O3_inline_sdc	5356	210,62
GCC49:adpcm_O3_vectorize_sdc	6135	110,81
GCC49:adpcm_Os_sdc	24397	68,45

Third example – Integer Division

Benchmark	CYCLES	HLS_execution_time
GCC49:dfdiv_none	1777	37,5
GCC49:dfdiv_nr1	1849	41,18
GCC49:dfdiv_nr2	1105	43,12
GCC49:dfdiv_NR	825	44,92
GCC49:dfdiv_as	841	30,14