CAOS tutorial





Access CAOS at: http://caos-iccd.necst.it/



Click on "click here to start" to start the optimization process





Page 2

Use CAOS to optimize your application

> Start a new project or load your work right were you left it





GUI Overview

CAOS Toolchain - NECST La	boratory			
WELCOME >> PROJE	CT >>	IR GENERATION		
IR Generation				Phase state: PROGRE
Input				
Code archive: Import	Progra	m description: Imp	ort 🗵	
CAOS module				
Hostname: m_1.1_from_doxygen	Port: 5011	up	date Status: ONLINE	
Output				
Run IR Generation				

Keep track of the **phase** that is being performed using the *navigation window*



GUI Overview

CAOS Toolchain - NECST Laboratory		
WELCOME >> PROJECT >> IR GENERATION		
IR Generation	Phase state:	PROGRESS
Input		
Code archive: Import Program description: Import		
CAOS module		
Hostname: m_1.1_from_doxygen Port: 5011 update Status: ONLINE		
Output		
Run IR Generation		

Check the status of the phase on the top-right part of the screen



GUI Overview



- > One of the main features of CAOS is **Modularity**
 - Every phase is backed up by a CAOS module that can be accessed simply specifying its hostname and port
 - The CAOS module panel provides information regarding the status of the specific module



CAOS Toolchain - NECST Laboratory	
WELCOME >> PROJECT >> IR GENERATION	
IR Generation	Phase state: PROGRESS
Input Code archive: Import Program description: Import	
CAOS module	
Hostname: m_1.1_from_doxygen Port: 5011 update Status: ONLINE	
Output	
Pup IP Concration	

- **IR Generation** generates the CAOS intermediate representation from the user's source code
- > For this phase you need to specify:
 - A zip/tar archive containing the source code of the application to optimize
 - A program description JSON file specifying the source code language and how to compile the application



Demo applications

$\dot{\leftarrow} \rightarrow \mathcal{C}$ \triangleq GitHub, Inc. [US] https://github.com/necst/tutorial_iccd17_code/ $\dot{\land}$ \Diamond \Diamond \bigcirc \bigcirc \bigcirc \bigcirc \bigcirc \bigcirc					
Branch: master - New pull re	equest	Create new file	Upload files	Find file	Clone or download -
lorenzoditucci removing un	used flags			Latest com	mit 2b69db6 5 hours ago
SDAccel	removing unused flags				5 hours ago
caos	updated caos vector sum application				a day ago
paper	adding paper for Smith-Waterman acceleration				6 hours ago
README.md	updated README				6 hours ago

- > Demo applications are available at: https://github.com/necst/tutorial_iccd17_code/
- > After downloading the repository, you can access the CAOS demo apps under the "caos" folder:
 - Vessel Segmentation
 - Smith-Waterman
- > In the following slides we will refer to the Vessel Segmentation demo application



Program Description



- Provides information regarding the source code:
 - Source code language
 - Supported compilers
 - Arguments needed for compiling



CAOS Toolchain - NECST Laboratory		
WELCOME >> PROJECT >> IR GENERATION		
IR Generation		Phase state: READY complete
Input		
Code archive: Import Program description: Import		
CAOS module		
Hostname: m_1.1_from_doxygen Port: 5011	update Status: ONLINE	
Output		
Run IR Generation		
Identified functions	Program static callgraph	
load_ppm		
main	main print matrix	
match_filter	IIIaIII print_iiiati1x	
print_matrix		
save_ppm		
	load_ppm match_filter save_ppm	

> Once the two files have been loaded, by clicking on **Run IR Generation** CAOS:

- Identifies all the functions within the application
- Provides a **Program Static Callgraph** highlighting the caller/calle relationships



CAOS Toolchain - NECST Laboratory		
WELCOME >> PROJECT >> IR GENERATION		
IR Generation		Phase state: READY complete
Input		
Code archive: Import 🛛 Program description: Import		
CAOS module		
Hostname: m_1.1_from_doxygen Port: 5011 upd	Status: ONLINE	
Output		
Run IR Generation		
Identified functions	Program static callgraph	
load_ppm		
main	main print matrix	
match_filter		
print_matrix		
save_ppm		
	load_ppm match_filter save_ppm	

Phase state becomes *READY* and we can click on *complete* to move to the next phase



CAOS Toolchain - NECST Laboratory					
WELCOME >> PROJECT >> IR GENERATION >>	APPLICABILITY CHECK				
IR Generation				Phase state: COMPLETED	edit
Input					
Code archive: Import Program description: Import					
CAOS module					
Hostname: m_1.1_from_doxygen Port: 5011	update Status: ONLINE				
Output					
Run IR Generation					
Identified functions	Program static callgra	uh sa			
load_ppm					
main		main	print matrix		
match_filter		IIIdIII	print_matrix		
print_matrix					
save_ppm					
	load_p	pm match_filter	save_ppm		

Phase state becomes *COMPLETED* and CAOS shows the new phase in the navigation window



APPLICABILITY CHECK

CAOS Toolchain - NECST Laboratory	
WELCOME >> PROJECT >> IR GENERATION >> APPLICABILITY CHECK	
Applicability Check	Phase state: PROGRESS
Input	
Architecture description: Import Architectural templates: Select	
CAOS module	
Hostname: m_1.2_default Port: 5012 update Status: ONLINE	
Output	
Run templates applicability check	

- APPLICABILITY CHECK verifies which CAOS architectural templates can be used for the given user code and target architecture
- > This phase requires:
 - A JSON file providing the target architecture description
 - A selection of the **architectural templates** to check





- Describes the target architecture on which the application will be executed
- > Two levels specification
 - Node description
 - System description

Let's define the architecture description for a system consisting of a single f1.2xlarge instance...



```
architecture-description.json
 1
 2
         "nodeDefinition" : {
 3
             "deviceTypes" : {
                  "f1-fpga" :
 4
                       "type" : "board",
 6
                      "vendor" : "Xilinx",
                      "partNumber" : "XCVU9P-FLGB2104-2-I"
 8
                  },
                  "intel-vcore" : {
10
                      "type" : "cpu",
                      "vendor" : "intel",
11
12
                      "partNumber" : "-"
13
             },
14
             "devices" : {
15
16
                  "f1-fpga-instance" : {
                      "type" : "f1-fpga"
17
18
                  },
                  "cpu" : {
19
20
                      "type" : "intel-vcore",
                      "host" : true
21
22
23
              },
24
              "connectionTypes" : { 🚥
31
             },
32
             "connections" : [ ----
38
39
         },
         "system" : { 🔤
40
44
45
         POLITECNICO MILANO 1863
                           POLITECNICO
         NECST
```

MILANO 1863

Page 15

EXTRA

- > Defines the type of devices to reference within the node specification
 - Intel CPU
 - Xilinx XCVU9P-FLGB2104-2-I board
- Instantiates the devices and specifies the device acting as "host" for the node



- Specifies the available connections between the devices of the node
- Defines how the devices are interconnected



DLITECNICO MILANO 186



Page 16



- Defines the number of nodes available within the system and how they are interconnected
- For the f1.2xlarge instance we simply define a single-node system



APPLICABILITY CHECK – Architectural Template

	Select architectural templates
Architectural templates: Select	Please select the architectural templates to consider: SST (version: 1.0) MasterSlave
5012 update Status: ONLINE	Maxeler Save

- > Select which architectural template(s) you want CAOS to consider for your application
 - Click Save
 - Click on Run templates applicability check



APPLICABILITY CHECK – Result

Architectural Templates Validation:

MasterSlave	
Status: Template is supported!	
Architecture report:	
Supported devices:	
• f1-fpga_instance	
Functions report:	
load ppm	
Hardware acceleration: no	
Reason: the function contains calls to unsup	oported functions
 Additional info: function: 'load_ppm' call: [for 	open]
main	
Hardware acceleration: no	
Reason: the function contains calls to unsur	oported functions
 Additional info: function: 'main' call: [printf] 	
match_filter	
Hardware acceleration: yes	
print matrix	
Hardware acceleration: no	
Reason: the function contains calls to unsu	oported functions
 Additional info: function: 'print_matrix' call: 	[printf]
save_ppm	
 Hardware acceleration: no 	
 Reason: the function contains calls to unsur 	apported functions
 Additional info: function: 'save_ppm' call: [fe 	open]
Mayolor	
Maxeler	

Status: Template not supported.

Reason: the system does not provide a supported device, or matches the specified connection rules Additional info: supported devices: board_Vectis

The Architectural Template Validation shows which architectural templates are being supported, and on which devices



APPLICABILITY CHECK – Result

Architectural Templates Validation:

MasterSlave		
Status: Template is supported!		
Architecture report:		
Supported devices:		
• f1-fpga_instance		
Functions report:		
load_ppm		
Hardware acceleration: no		
Reason: the function contains calls to unsupported functions		
Additional info: function: 'load_ppm' call: [fopen]		
main		
Hardware acceleration: no		
Reason: the function contains calls to unsupported functions		
Additional info: function: 'main' call: Invintf		
match_filter		
Hardware acceleration: yes		
Hardware acceleration: no		
Reason: the function contains calls to unsupported functions		
Additional info: function: 'print_matrix' call: [printf]		
Hardware acceleration: no		
Reason: the function contains calls to unsupported functions		
Additional info: function: 'save_ppm' call: [fopen]		
Maxeler		
Status: Template not supported		

Reason: the system does not provide a supported device, or matches the specified connection rules Additional info: supported devices: board_Vectis

> Provides information on which functions are **suitable** for a **hardware acceleration**



PROFILING

CAOS Toolchain - NECST Laboratory				
WELCOME >> PROJECT >> IR GENERATION >> APPLICABILITY	CHECK >> PROFILING			
Profiling	Import Dataset			
Input Profiling dataset: Add 🗵	Dataset name:			
CAOS module Hostname: m_1.3_cpp_perf_prof Port: 5013 update Status: ONLINE Output	Please select a dataset archive: Scegli file Nessun file selezionato Command line arguments:			
Profile datasets	NOTE: use %%DATASET_DIR%% to refer to the directory where the archive will be extracted SUBMIT			

- > Click "Add..." on *Profiling Dataset* to provide a dataset for the profiling phase:
 - Choose a *Name*
 - Upload an Archive containing the dataset (input files for the application)
 - Command line arguments necessary for the execution
- > N.B. If the source code auto-generates the dataset, the archive and arguments might not be needed



PROFILING

CAOS Toolchain - NECST Laboratory										
WELCOME >> PROJECT >> IR GENERATION >> APPLICABILITY	CHECK >> PROFILING									
Profiling	Import Dataset									
Input Profiling dataset: Add 🗵	Dataset name:									
CAOS module Hostname: m_1.3_cpp_perf_prof Port: 5013 update Status: ONLINE Output	Please select a dataset archive: Scegli file Nessun file selezionato Command line arguments:									
Profile datasets	NOTE: use %%DATASET_DIR%% to refer to the directory where the archive will be extracted SUBMIT									

Click SUBMIT and Profile datasets to start code profiling





Output

Profile datasets

Dataset: NO_DATASET

Function	Self time %	Total time %
match_filter(unsigned char[1080][1440], unsigned char[1080][1440])	98.96%	98.96%
main(int, char *[])	0.21%	99.25%
Overall external calls	0.84%	n.a.

> Profiling result

- Identifies the most **computationally intensive** functions of the application
- Provides percentages regarding self and total execution time



PARTITIONING

CAOS Toolchai	AOS Toolchain - NECST Laboratory										
WELCOME	>>	PROJECT	>>	IR GENERATION	>>	APPLICABILITY CHECK	>>	PROFILING	>>	PARTITIONING	
HW / SW partitioning											
CAOS module	CAOS module										
Hostname: m_1.	4_default	Por	t: 5014	up	date	Status: ONLINE					
Output											

Run partitioning

Function			Profilir	ng Data	HW / SW partitioning		
		Self time %			otal time	%	HW acceleration per architectural template
		avg	max	min	avg	max	MasterSalve
match_filter(unsigned char[1080][1440], unsigned char[1080][1440])	98.96%	98.96%	98.96%	98.96%	98.96%	98.96%	- let partitioner decide - 🖨
main(int, char *[])	0.21%	0.21%	0.21%	99.25%	99.25%	99.25%	Software
Overall external calls	0.84%	0.84%	0.84%	n.a.	n.a.	n.a.	

For each function that is candidate for Hardware Acceleration it is possible to decide to force a hardware or software implementation, or to let CAOS decide





HW / SW partitioning

CAOS module

Output

Run partitioning

			Profilir	ng Data	HW / SW partitioning		
Function	Self time %			Total time %			HW acceleration per architectural template
		avg	max	min	avg	max	MasterSalve
match_filter(unsigned char[1080][1440], unsigned char[1080][1440])	98.96%	98.96%	98.96%	98.96%	98.96%	98.96%	Hardware 🛟
main(int, char *[])	0.21%	0.21%	0.21%	99.25%	99.25%	99.25%	Software
Overall external calls	0.84%	0.84%	0.84%	n.a.	n.a.	n.a.	

After clicking Run partitioning CAOS automatically selects which functions to accelerate on hardware



FUNCTIONS OPTIMIZATION

CAOS Toolchain - NECST Laboratory	
WELCOME >> PROJECT >> IR GENERATION >> APPLICABILITY CHECK >> PROFILING >> PARTITIONING >> FUNCTIONS OPTIMIZATION	
Functions optimization	Phase state: READY complete
Code archives	
Architectural template: MasterSalve	
Initial Code 9	
Start optimization	

Within the Functions Optimization phase, we can specify the version of the code we want to work on by clicking on the corresponding folder

Page 26

Initially, only one version of the code is available

POLITECNICO

WILANO 1863

After selecting the code version, we can either decide to:
 a) Click *complete* and move to the next phase
 b) Click on *Start Optimization* to start optimizing the code



FUNCTIONS OPTIMIZATION – Pre-Opt IR Generation

Functions optimization

CANCEL	PRE-OPT IR GENERATION							
CAOS module	le							
Hostname: m_1	_1.1_from_doxygen Port: 5011 update Status: ONLINE							
Output								
Generate IR								
Identified functi	tions							
load_ppm(cha	ar *, unsigned char[1080][1440], int)							
main(int, char	r *[])							
match_filter(un	match_filter(unsigned char[1080][1440], unsigned char[1080][1440])							
print_matrix(sh	short[1080][1440])							
save_ppm(cha	nar *, unsigned char[1080][1440])							

- > During **Function Optimization** there are multiple sub-phases
- > This first one, generates an *intermediate representation* of the code before optimizing it



FUNCTIONS OPTIMIZATION – Static Code Analysis

Functions optimization

SENERATION >>	STATIC CODE ANA	LYSIS								
CAOS module										
Port: 5021	update Statu	IS: ONLINE								
		Property	Value							
		averageLatency	7002323881							
match_filter(unsigned char[1080][1440], unsigned char[1080][1440]										
		worstLatency	7002323881							
	SENERATION >> Port: 5021 080][1440], unsign	SENERATION >> Port: 5021 update Statu 080][1440], unsigned char[1080][1440]	SENERATION >> Port: 5021 update Status: Status: ONLINE Property averageLatency bestLatency worstLatency worstLatency							

The Static Code Analysis phase provides a set of properties (such as expected latency) for all the functions that are candidates for hardware acceleration



FUNCTIONS OPTIMIZATION – Hardware Estimation

	Functi	ons optimization	Phase state: Pf	ROGRESS					
	CANCEL	PRE-OPT IR GENERATION	>> STAT	IC CODE ANAI	YSIS	>> HARDWAR	E ESTIMATION		
CA	OS module							Phase state: READY	complete
н	ostname: m_2.	2_default Port: 5022		update Statu	s: ONLIN	E			
Ha	Run hardware (estimation							
F	unction				Device	Resource Type	Quantity		
						BRAM_18K	3 (0.07%)		
						DSP48E	7 (0.10%)		
n	match_filter(unsigned char[1080][1440], unsigned char[1080][1440])			[1080][1440])	f1-fpga	FF	1024 (0.04%)		
					LUT	1597 (0.14%)			
						URAM	0 (0.00%)		

By clicking on Run hardware estimation CAOS estimates the amount of resources needed for implementing the functions on the FPGA



FUNCTIONS OPTIMIZATION – Performance Estimation

Functions optimization Phase state: PROGRESS **PERFORMANCE ESTIMATION** CANCEL STATIC CODE ANALYSIS HARDWARE ESTIMATION PRE-OPT IR GENERATION >> >> >> **CAOS** module Phase state: READY complete Hostname: m 2.3 default Port: 5023 update Status: ONLINE Run performance estimation Functions performance estimation and suggested optimization Top function: match_filter(unsigned char[1080][1440], unsigned char[1080][1440]) **Current function implementation Optimization 1) Optimization 2)** Performance estimation Optimization type: on chip full caching **Optimization type:** pipelining Execution time: 35.012 s After hardware estimation, Performance estimation Performance estimation Clock frequency: 200 MHz Execution time: 29.056 s Execution time: 13.750 s CAOS estimates Hardware estimation Clock frequency: 200 MHz Clock frequency: 200 MHz BRAM 18K 3 (0.07%) Hardware estimation Hardware estimation performance suggesting DSP48E 7 (0.10%) BRAM_18K 1523 (**35.25%**) BRAM_18K 3 (0.07%) potential optimizations f1-fpga FF 1024 (**0.04%**) DSP48E DSP48E 7 (0.10%) 7 (0.10%) LUT 1597 (**0.14%**) f1-fpga FF f1-fpga FF 1211 (0.05%) 1274 (0.05%) URAM 0 (0.00%) LUT 2211 (0.19%) LUT 1937 (0.16%) URAM 0 (0.00%) URAM 0 (0.00%) Parameters: Parameters: fpga_to_host_copy: [0,1] II: host to fpga copy: [0,1] relativeForLine: 15 POLITECNICO

Page 30

MILANO 1863

EXTRA

FUNCTIONS OPTIMIZATION – Code Optimization

Functions optimization					Phase state: PROGRESS
CANCEL PRE-OPT IR GENERATION >> STATIC CODE ANAI	LYSIS >> HARDWA	ARE ESTIMATION >>	PERFORMANCE ESTIMATION	>> CODE OPTIMIZATION	
CAOS module				Phase	e state: READY complete
Hostname: m_2.4_default Port: 5024 update Statu	s: ONLINE				
Input					
Function Optimiza	tion				
match_filter(unsigned char[1080][1440], unsigned char[1080][1440]) Optimization	on 1 (on_chip_full_caching) ᅌ	J			
Output					
Apply code optimizations					
Optimization Report					
function	optimization applied				
load_ppm(char *, unsigned char[1080][1440], int)	NO				
main(int, char *[])	NO				
match_filter(unsigned char[1080][1440], unsigned char[1080][1440])	YES				
print_matrix(short[1080][1440])	NO				
save_ppm(char *, unsigned char[1080][1440])	NO				

- > Within Code Optimization sub-phase, it is possible to select among the suggested optimizations
- > By clicking *Apply code optimizations* the code will be modified accordingly



FUNCTIONS OPTIMIZATION – Post-Opt IR Generation

	CANCEL	PRE-OPT IR G	ENERATION	>>	STATIC CODE	E ANALYSIS	>>		
	POST-OPT II	R GENERATION							
C	AOS module	•							
	Hostname: m_1	.1_from_doxygen	Port: 5011		update	Status: ONL	NE		
	Output Regenerate IR								
10		ons • * ······	40001[4.4.40] ::::	LN					
	load_ppm(cnar	*, unsigned char[1080j[1440], in	[)					
	main(int, char *	·[])							
	match_filter(unsigned char[1080][1440], unsigned char[1080][1440])								
	print_matrix(sh	ort[1080][1440])							
	save_ppm(cha	r *, unsigned char[1080][1440])						

- > After having applied one of the suggested code optimizations the IR might have changed
- > The Post-Opt IR Generation sub-phase allows to regenerate the IR.



FUNCTIONS OPTIMIZATION



- > The optimization process is complete, we can either:
 - try to apply more optimizations by selecting the new code version and clicking on Start Optimization
 - select the desired code version and move to the implementation phase



IMPLEMENTATION

CAOS Toolchain - NECST Laboratory													
WELCOME >>	PROJECT	>>	IR GENERATION	>>	APPLICABILITY CHECK	>>	PROFILING	>>	PARTITIONING	>>	FUNCTIONS OPTIMIZATION	>>	IMPLEMENTATION
Implemen	tation											Phas	e state: READY complete
CAOS module													
Hostname: m_3_sdacc	el Po	rt: 5030	up	odate	Status: ONLINE								
Run implementation													
Implementation arch	ive												

- > During the Implementation phase CAOS will produce the host, kernel file and Makefile for SDAccel
- > Once the phase is completed, click on the Implementation Archive to download the code
- > After running the implementation, the output archive is also accessible on the server at:

~/Documents/CAOS_outputs



Emulate / Build the final application

- > Upload the CAOS archive to an AWS instance
 - scp -i <pem file> <CAOS-ARCHIVE>.zip centos@<public_dns entry>:~/
- > SSH to the AWS Instance and load the SDAccel settings
 - -ssh -i <pem file> centos@<public_dns entry>
 - cd \$AWS_FPGA_REPO_DIR
 - source sdaccel_setup.sh
 - source /opt/Xilinx/SDx/2017.1.op/settings64.sh

> Unzip the CAOS archive

- -cd ~/
- unzip **<CAOS-ARCHIVE>.**zip
- -cd output

> Use the Makefile to run HW / SW emulation or build the application

- make emulation TARGET=sw_emu
- make emulation TARGET=hw_emu
- make build TARGET=hw



Save the CAOS project

CAOS Toolchain - NECST Laboratory										
WELCOME >> PROJECT >> IR GENERATI	ION >> APPLICABILITY CHEC	CK >> PROFILING	>> PARTITIONING >>	FUNCTIONS OPTIMIZATION >>	IMPLEMENTATION					
Manage your CAOS project ⑦ Save current project Close current project)									

> CAOS flow is over, remember to save your project for future use, and then close it.

