Design for Low-Power IoT Systems:
Coarse-Grained Reconfigurable Acceleration Units

FRANCESCA PALUMBO

UNIVERSITÀ DEGLI STUDI DI SASSARI

Design for Low-Power IoT Systems

Module 1: Transport Triggering Approach
Jarmo Takala, Tampere University of Technology, Finland

Module 2: Coarse-Grained Reconfigurable Acceleration Units
Francesca Palumbo, Università degli Studi di Sassari, Italy

Module 3: Markov Decision Processes and Power/Performance/Thermal (PPT)
Marilyn Wolf, Georgia Tech., USA

Module 4: Factored MDPs and Application Examples
Shuvra Bhattacharyya, U. Maryland, USA and Tampere U. Technology, Finland

Overview

- Motivations
  - What we need adaptation for
  - Triggers and Types
- Coarse-Grained Reconfigurable Systems
  - Computing Spectrum and Reconfigurable Systems Classification
  - Heterogeneous and Irregular Coarse-Grained Reconfigurable Accelerators
- Power Management
  - Issues and Strategies
  - Low-Power Coarse-Grained Reconfigurable Accelerators
- Reconfigurable FFT Example and Remarks

Overview

- Motivations
  - What we need adaptation for
  - Triggers and Types
- Coarse-Grained Reconfigurable Systems
  - Computing Spectrum and Reconfigurable Systems Classification
  - Heterogeneous and Irregular Coarse-Grained Reconfigurable Accelerators
- Power Management
  - Issues and Strategies
  - Low-Power Coarse-Grained Reconfigurable Accelerators
- Reconfigurable FFT Example and Remarks

Numbers: opportunity or issue?

> 7 billion

20 MWh/year

=

1,800 kg oil

Designed by Freepik

> 1 billion smartphones

8.4 billion connected things in 2017 (+31% wrt 2016)

20.4 billion by 2020

http://www.gartner.com/newsroom/id/3598917

Some examples …

Connectivity and real-time situation awareness are nowadays common in different scenarios.

SMART HEALTH:
distributed healthcare assistance to improve quality of life and active and healthy ageing, functionalities can be changed according to the daily tasks.

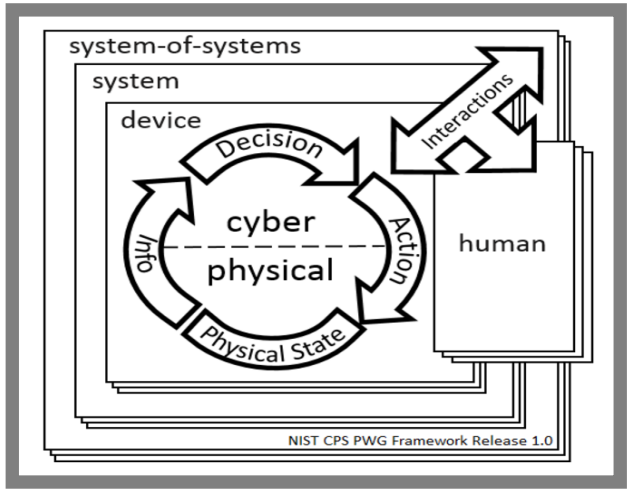
SMART-SOCIETY:
increased building efficiency and comfort, i.e. lightning/air quality management can be adjusted to the room status.

SMART-TRANSPORTATION:
autonomous electric vehicle, improved driver assistance and care. Path towards destinations may vary, even diverging from the optimal one, according to user preferences.

Reconfiguration: Recipe for Compromises

Modern digital devices (*real-time* and *ad-hoc*) are pervasive (*98%* of computers are *embedded*) and interconnected.

They may also present *sensing* and *actuating capabilities*, leading to the concept of CPS.



	Safety	Security	Certif.	Distrib.	HMI	Seamless	MPSoC	Energy
Automotive	x	x	x	x	x	x	x	
Aerospace	x	x	x	x	x		x	x
Healthcare	x	x	x	x	x	x	x	x
Consumer					x	x	x	

Reconfiguration may allow to optimally implement complex/demanding systems, managing ***numerous/conflicting requirements*** and a ***variety of functionalities***.

Triggers for Adaptation

ENVIRONMENTAL AWARENESS:

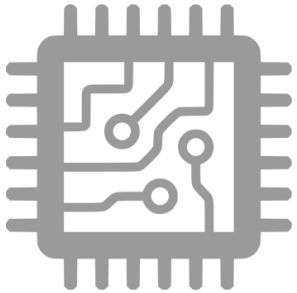
Influence of the environment on the system, i.e. daylight vs. nocturnal, radiation level changes, etc.

Sensors are needed to interact with the environment and capture conditions variations.

USER-COMMANDED:

System-User interaction, i.e. user preferences, etc.

Proper human-machine interfaces are needed to enable interaction and capture commands.

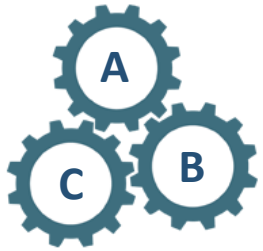


SELF-AWARENESS:

The internal status of the system varies while operating and may lead to reconfiguration needs, i.e. chip temperature variation, low battery.

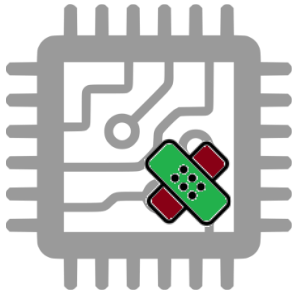
Status monitors are needed to capture the status of the system.

Types of Adaptation



FUNCTIONALITY-ORIENTED:
To adapt functionality because the CPS mission changes, or the data being processed changes and adaptation is required.
It may be parametric (a constant changes) or fully functional (algorithm changes).

NON-FUNCTIONAL REQUIREMENTS-ORIENTED:
Functionality is fixed, but system requires adaptation to accommodate to changing requirements, i.e. execution time or energy consumption.

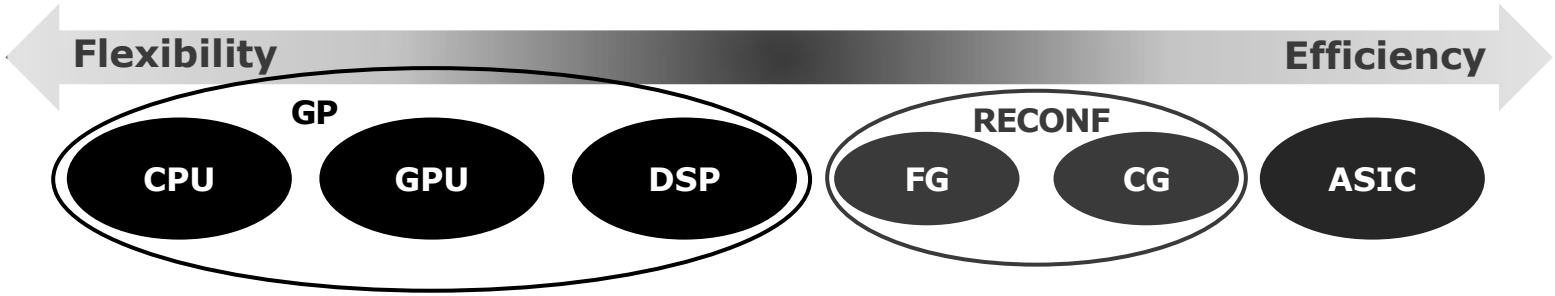


REPAIR-ORIENTED:
For safety and reliability purposes, adaptation may be used in case of faults. Adaptation may add self-healing or self-repair features. e.g.: HW task migration for permanent faults, or scrubbing (continuous fault verification) and repair.

Overview

- Motivations
  - What we need adaptation for
  - Triggers and Types
- Coarse-Grained Reconfigurable Systems
  - Computing Spectrum and Reconfigurable Systems Classification
  - Heterogeneous and Irregular Coarse-Grained Reconfigurable Accelerators
- Power Management
  - Issues and Strategies
  - Low-Power Coarse-Grained Reconfigurable Accelerators
- Reconfigurable FFT Example and Remarks

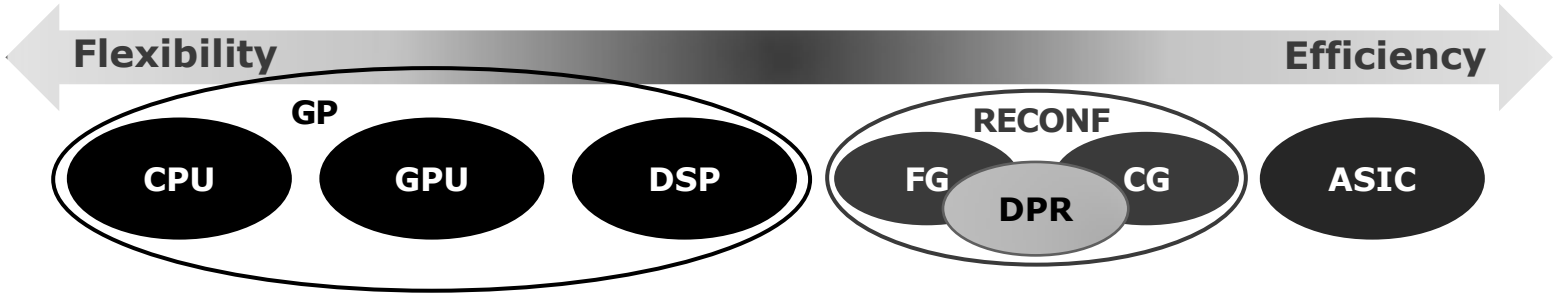
Computing Spectrum



Reconfigurable computing provides a **trade-off** between execution **efficiency** typical of ASICs and **flexibility** mainly exhibited by GP devices.

	Fine-Grained (FG) bit-level	Coarse-Grained (CG) word-level
flexibility	☺	😐
speed	☹	☺
memory	☹	☺

Computing Spectrum



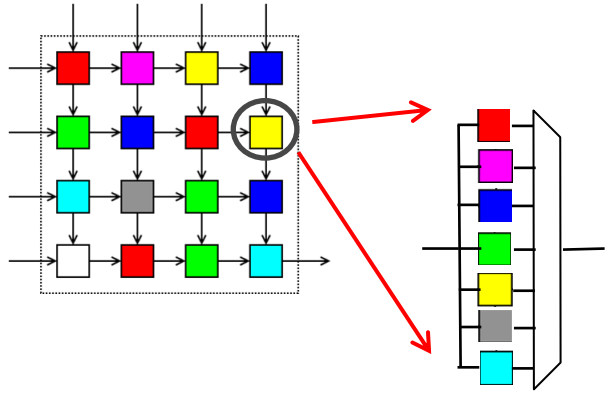
Reconfigurable computing provides a **trade-off** between execution **efficiency** typical of ASICs and **flexibility** mainly exhibited by GP devices.

	Fine-Grained (FG) bit-level	Dynamic Partial (DP) bit-level	Coarse-Grained (CG) word-level
flexibility	☺	😐	😐
speed	☹	😐	☺
memory	☹	😐	☺

Virtual vs. Dynamic & Partial

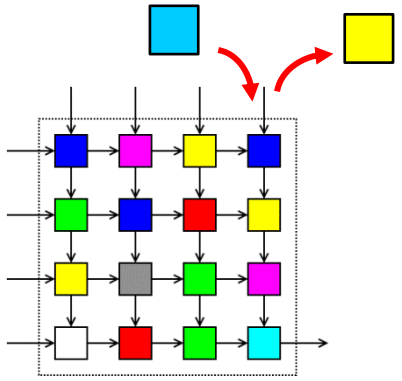
VRC → Virtual Reconfigurable Circuits

- High reconfiguration speed
- Lower operation speed (mux and size)
- Higher Area Overhead
- Technology independent (ASIC or FPGA)



DPR → Dynamic and Partial Reconfiguration

- Lower reconfiguration speeds
- Better operation speed (no mux/less logic)
- Better Resource Utilization (no dark logic)
- Higher Flexibility and Scalability
- Technology dependent (FPGA)



Multi-Dataflow Composer: Development

2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018

Baseline tool specification:
Multi-Dataflow Composer (MDC) tool

MPEG-RVC Framework Integration:
Orcc + MDC + Xronos + Turnus

Structural Profiler

Low-Power Extension

Co-processor Generator

Generalization Process

Multi-Dataflow Composer: Evaluation

2010 2011 2012 2013 2014 2015 2016 2017 2018

Reconfigurable Image/Video Coding: JPEG e H.264

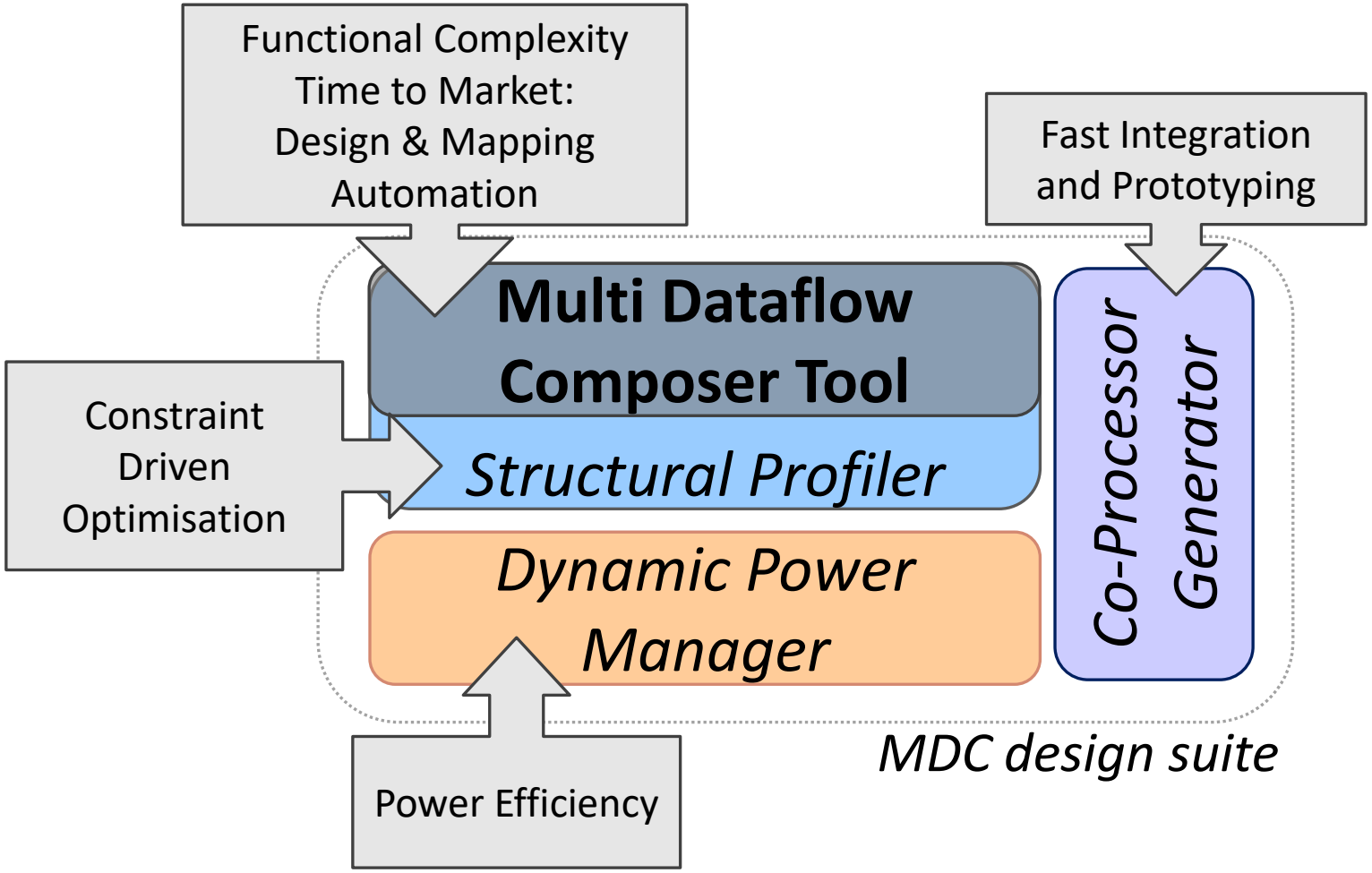
Neural Signal Decoding

Adaptive Filtering: HEVC Encoding

Cryptographic Systems

CERBERO H2020

Multi-Dataflow Composer: Design Suite

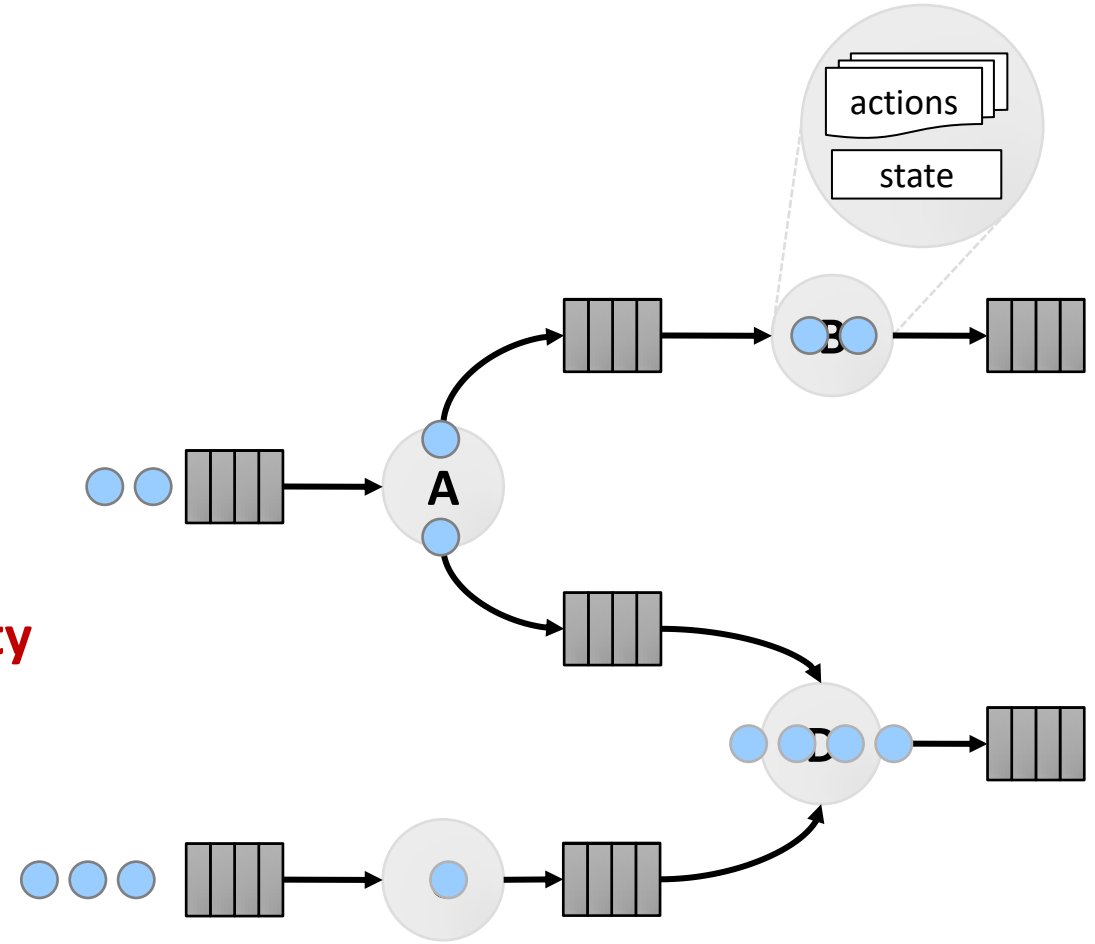


http://sites.unica.it/rpct/

Dataflow Model of Computation

COMPUTING PARADIGM:

- directed flow graph of **actors** (functional units)
- communication with **tokens** (packets of data) exchange through dedicated channels



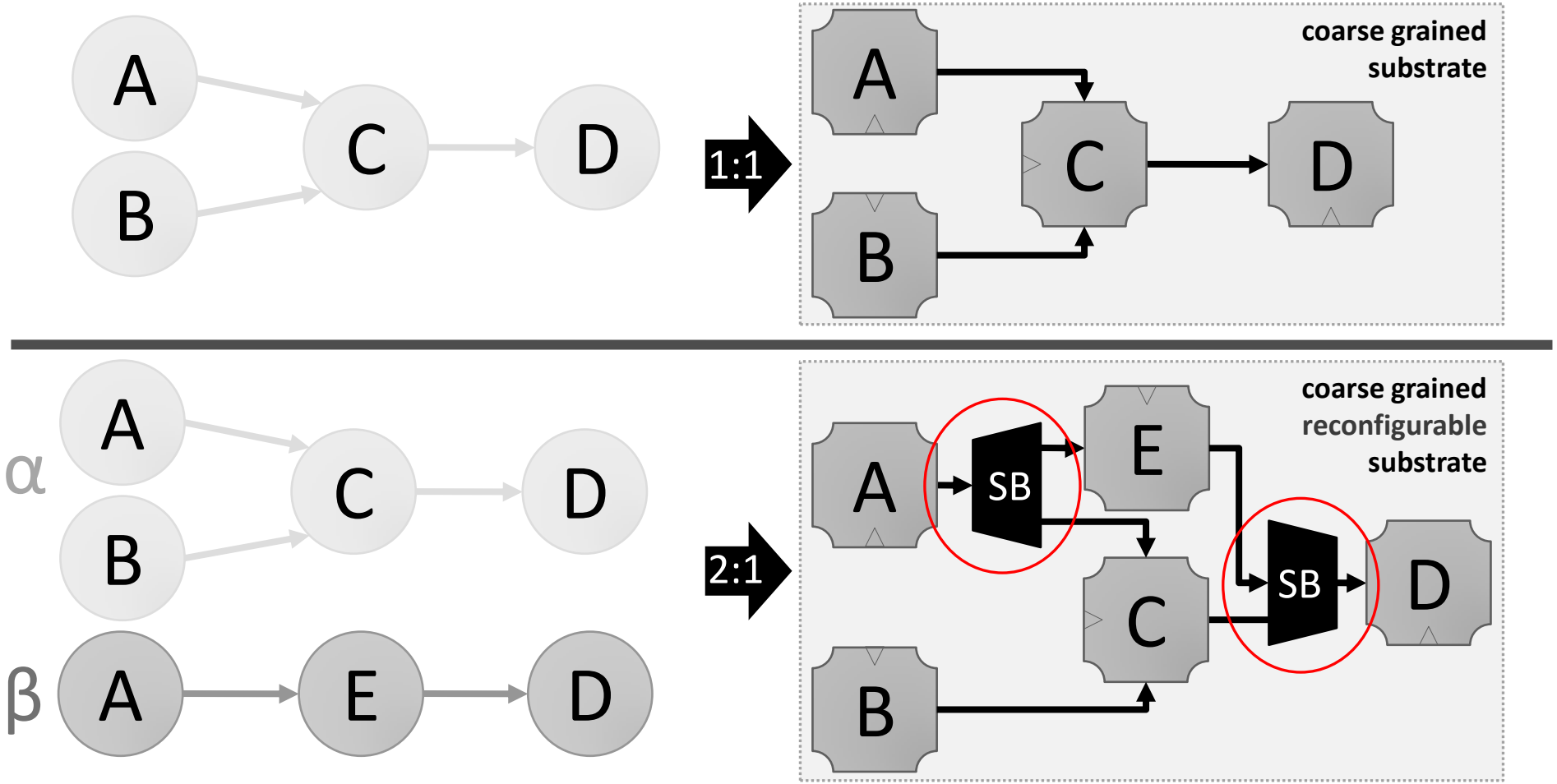
PECULIARITIES:

- explicit intrinsic application **parallelism**
- modularity favours model **re-usability/adaptivity**

EXTERNAL INTERFACE:

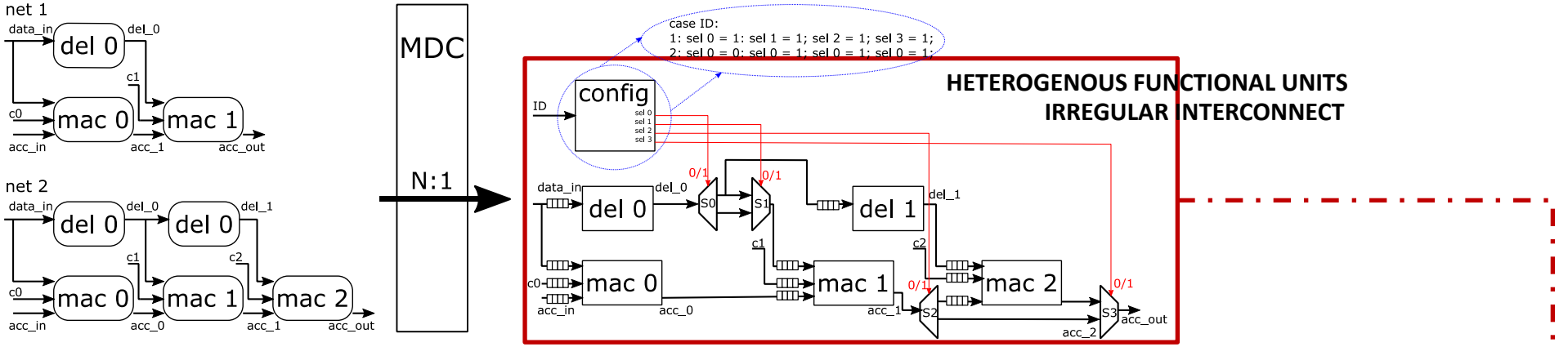
- I/O **ports number**
- I/O **ports depth**
- I/O ports tokens **burst**

Heterogeneous and Irregular: approach



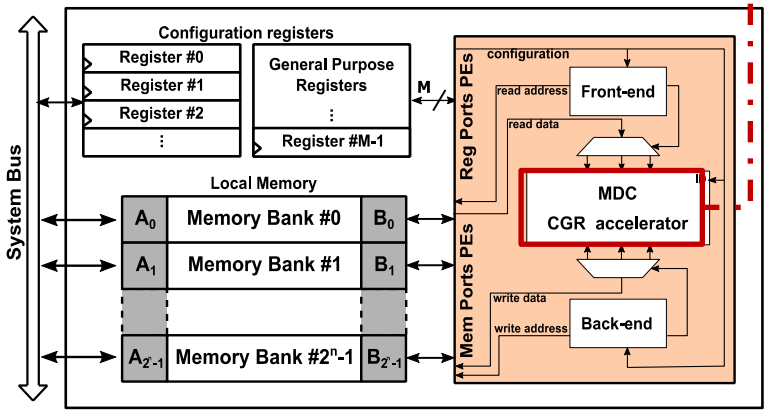
http://sites.unica.it/rpct/

Heterogeneous and Irregular: MDC framework



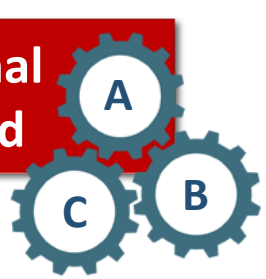
Multi-Dataflow Composer (MDC) tool: Dataflow 2 HW tool

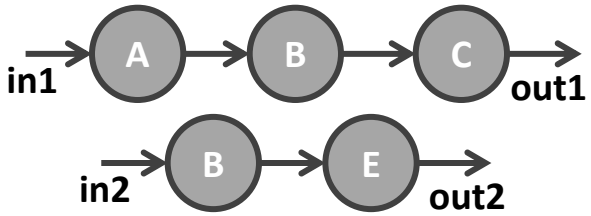
- Given N input dataflow specification, it generates the HDL Coarse-Grain (CG) reconfigurable substrate
- Handles programmability, by defining switching and configuration logic
- Deploy Xilinx-compliant IP blocks, plus their drivers

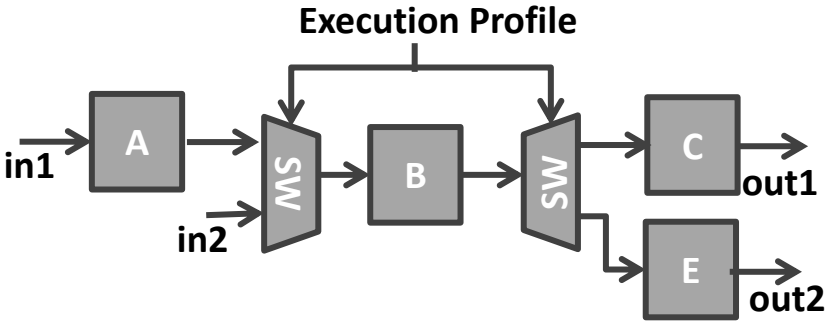


MDC-based Reconfiguration: Adaptation Types

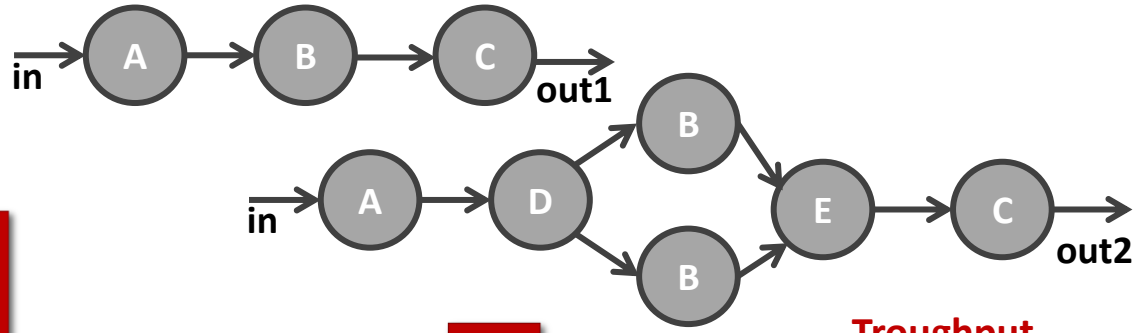
Functional Oriented

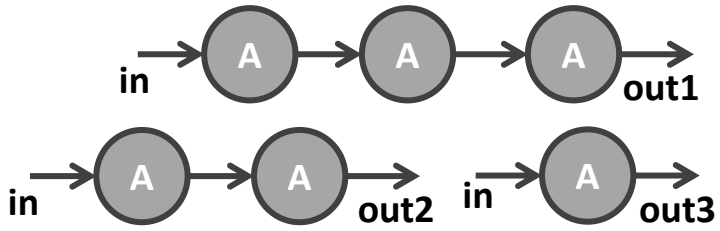




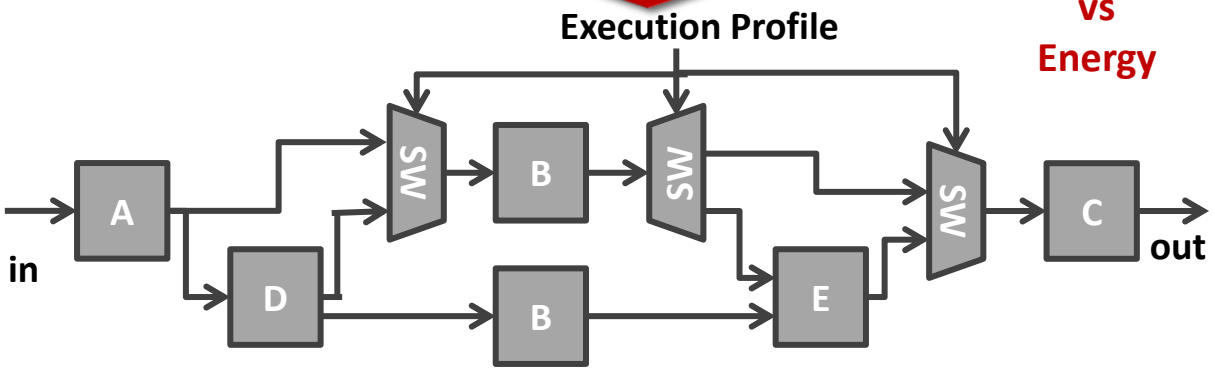


Non-Functional Oriented

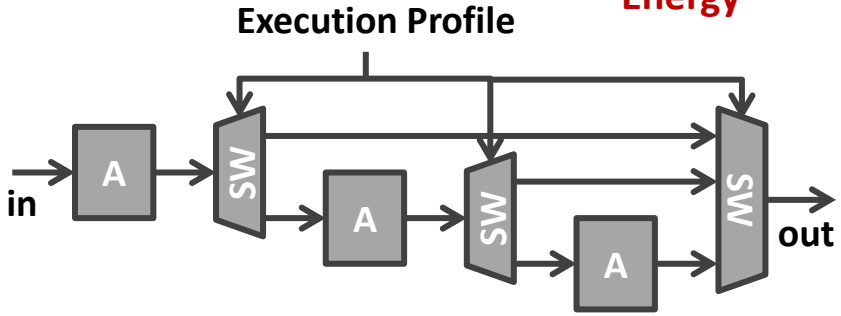




Troughput vs Energy



QoS vs Energy



Overview

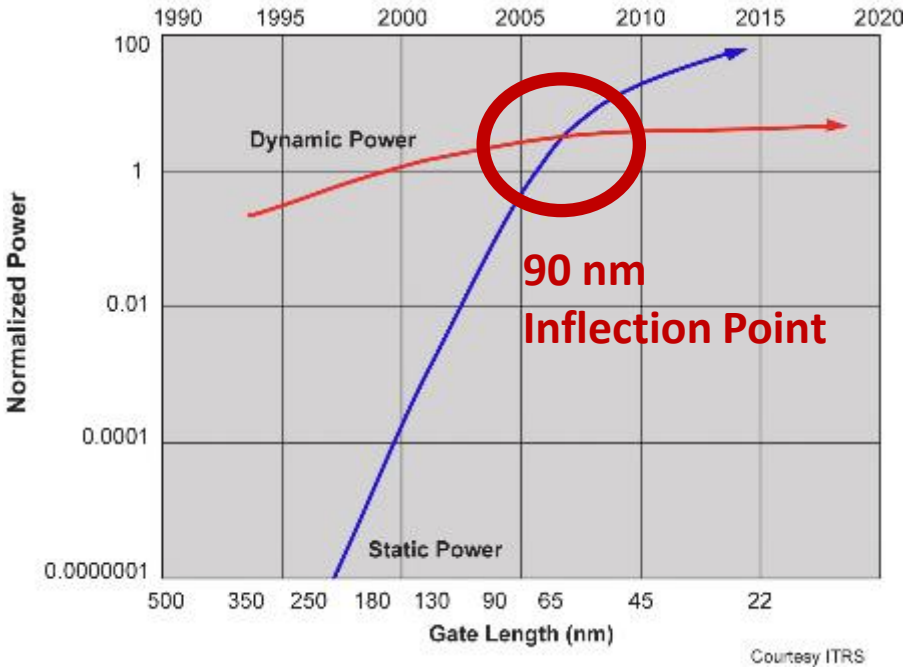
- Motivations
  - What we need adaptation for
  - Triggers and Types
- Coarse-Grained Reconfigurable Systems
  - Computing Spectrum and Reconfigurable Systems Classification
  - Heterogeneous and Irregular Coarse-Grained Reconfigurable Accelerators
- Power Management
  - Issues and Strategies
  - Low-Power Coarse-Grained Reconfigurable Accelerators
- Reconfigurable FFT Example and Remarks

The Power Issue

Power consumption =
Dynamic power + Static power

Dynamic: activity dependent

- *Short-circuit*: when the output line of a transistor is switching, there is a period of time when both the PMOS and the NMOS transistors are on (I·V·f)
- *Switching power*: due to the charging and discharging of the load capacitance when logic transitions occur (determined by the formula C·V2·f).



Static: not activity dependent, but due to leakage currents.

- Do not depend on switching and operating frequency.
- As transistors get smaller, channel lengths become shorter and leakage currents increase.

The Power Issue: Textbook Techniques

Dynamic	***Static***
Clock gating	Multi-threshold dev.
Variable frequency	Power gating
Voltage islands	Back (substrate) bias
Variable/multi power supply	Multi-oxide devices
Dynamic Voltage & Frequency Scaling	SOI CMOS

- ***DESIGN***: Multi Vt, Clock gating, Power gating, Multi Vdd, DVFS.
- ***PROCESS***: Multi Vt, PD SOI, FD SOI, FinFet, Body Bias, Multi oxide devices, Minimize capacitance by custom design.

MANAGEABLE AT SYSTEM LEVEL

- ***ARCHITECTURE***: power-constrained DSE, hw customization and IP specific techniques, parallelism, mapping.

INTRINSICALLY SYSTEM LEVEL

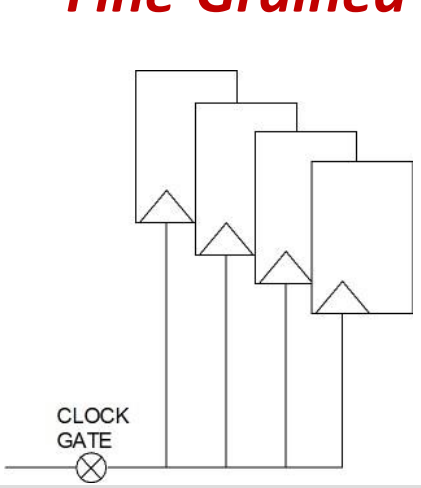
Clock-Based Techniques

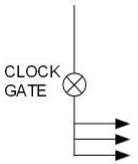
Reduce frequency whenever you can.
Stop the clock when the component is not active.

- Power consumed by flip-flops.
- Power consumed by combinatorial logic driven by registers.
- Power consumed by the clock buffer tree.

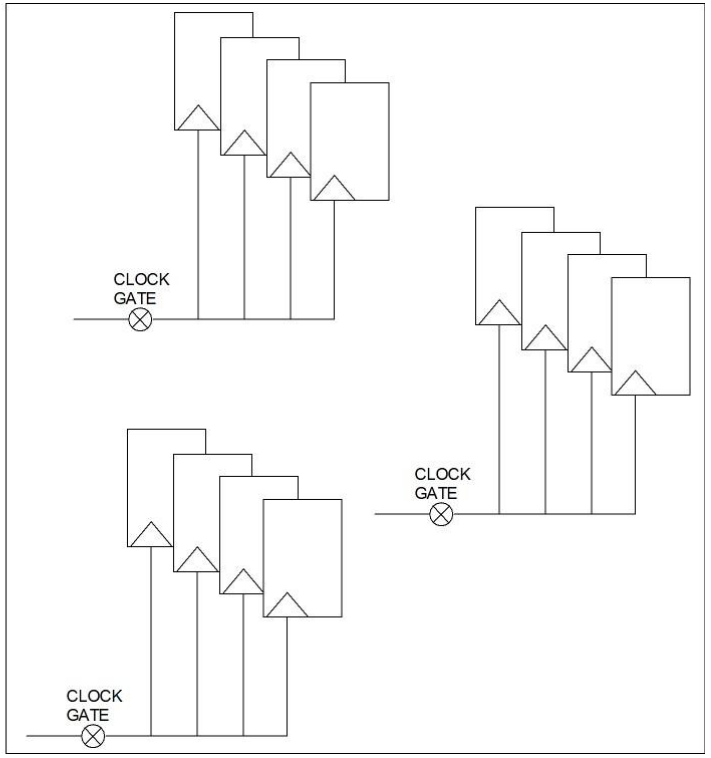
50% less dynamic power

Fine-Grained





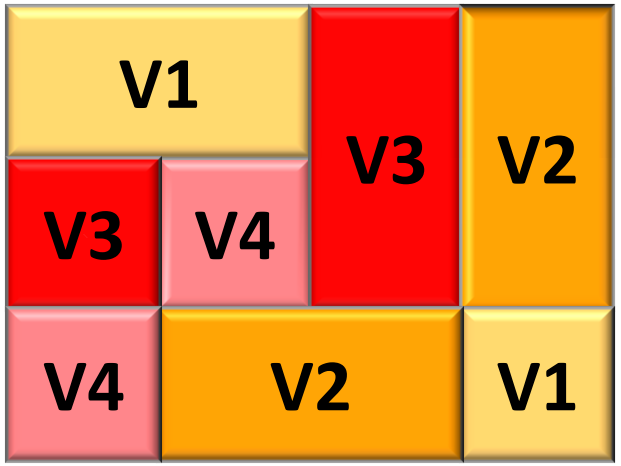
Coarse-Grained



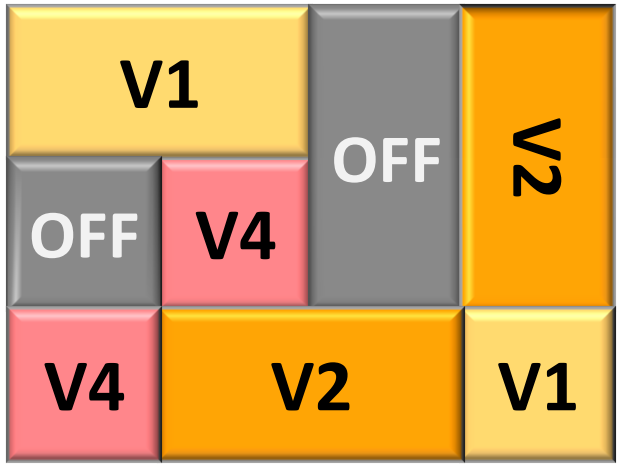
Power-Based Techniques

Reduce the voltage level of a power island whenever you can.
Switch it off when it is not active.

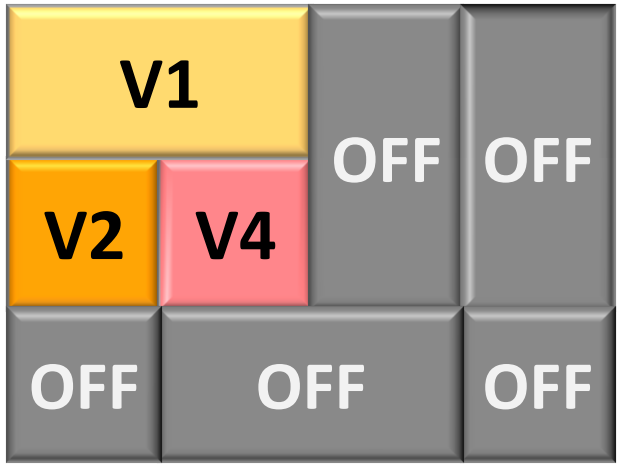
- Power-aware partitioning.
- Switching-off power island brings local leakage to zero.
- Modes of operation -> power down all the idle chip regions.



Static Voltage Scaling (SVS)



SVS with Power Shut Off (PSO)



SVS with PSO and Power Modes

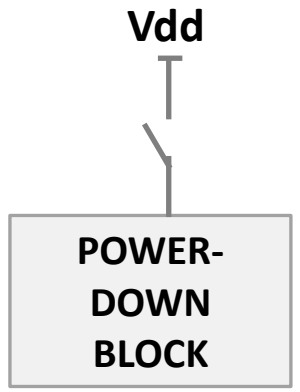
Clock-/Power-Based Techniques: Overhead

Switch off the clock, applicable to both ASIC and FPGA.

- @ASIC: simple and gates.
- @FPGA: dedicated vendor specific cells.

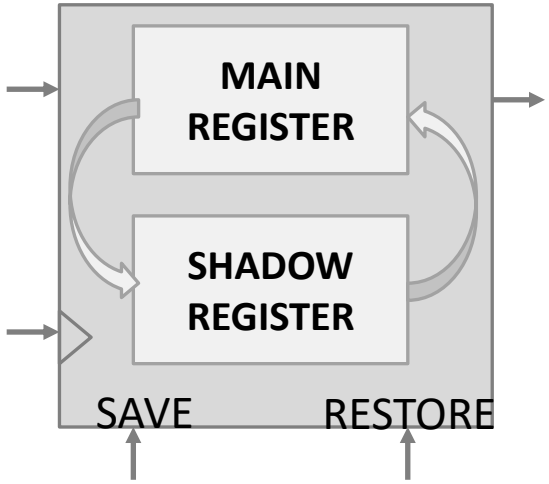
Switch off the power supply, ASIC only (partially FPGA).

- *Sleep transistors*: to switch on and off power supply.
- *Isolation logic*: to avoid the transmission of spurious signals from gated regions to normally-on cells.
- *Retention logic*: to maintain the internal state of gated regions.

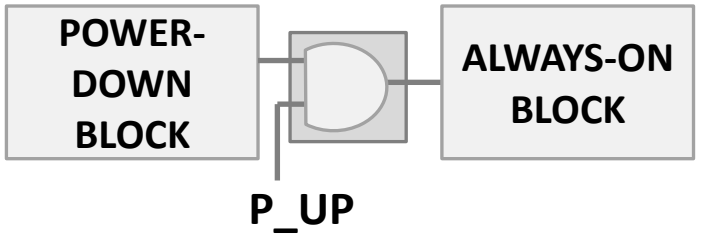


Power Switch-Off Cell

Retention Register



Isolation Cell



Clock-/Power-Based Techniques: Overhead

Switch off the clock, applicable to both ASIC and FPGA.

- @ASIC: simple and gates.
- @FPGA: dedicated vendor specific cells.

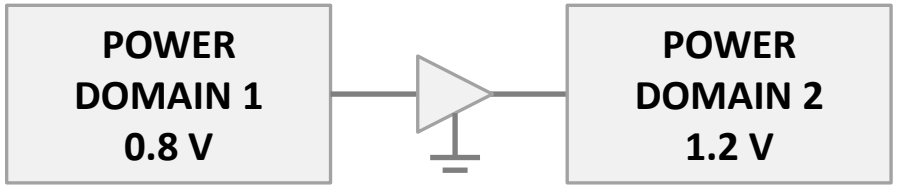
Switch off the power supply, ASIC only (partially FPGA).

- *Sleep transistors*: to switch on and off power supply.
- *Isolation logic*: to avoid the transmission of spurious signals from gated regions to normally-on cells.
- *Retention logic*: to maintain the internal state of gated regions.

Vary mode changing Vdd, ASIC only.

- *Level shifters*: to pass signals between portions of the design that operate on different voltages.

Level Shifters

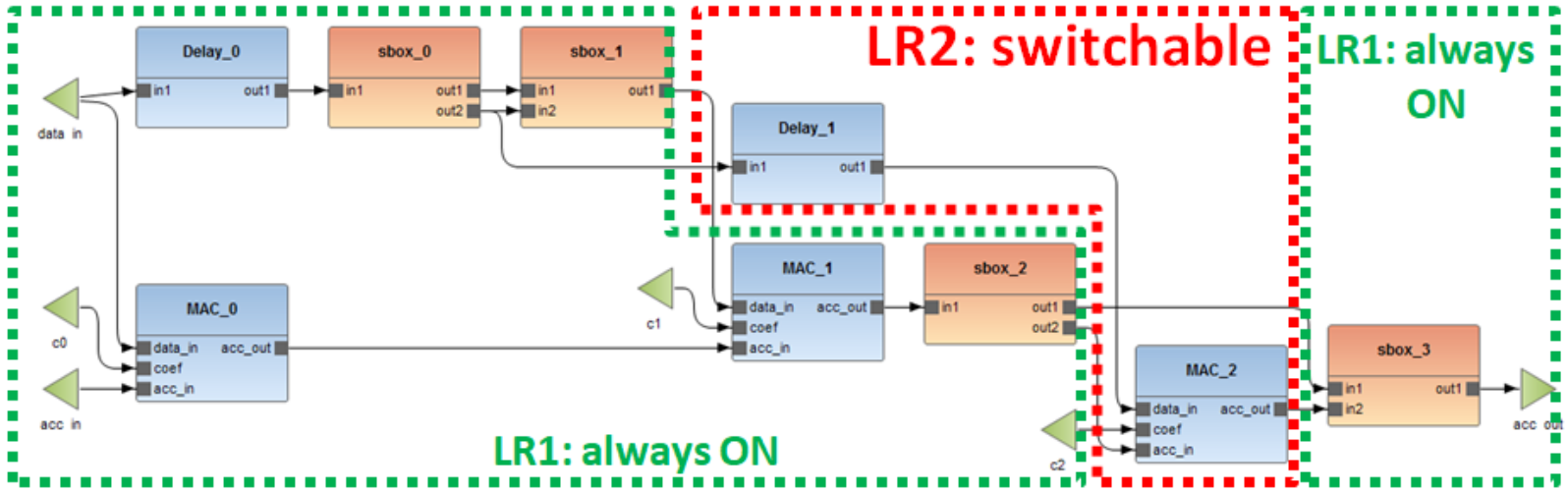


Clock-/Power-Based Techniques: Management

The ***Common Power Format (CPF)***, used within the Cadence design flow, is meant to define power-saving techniques early in the design process.

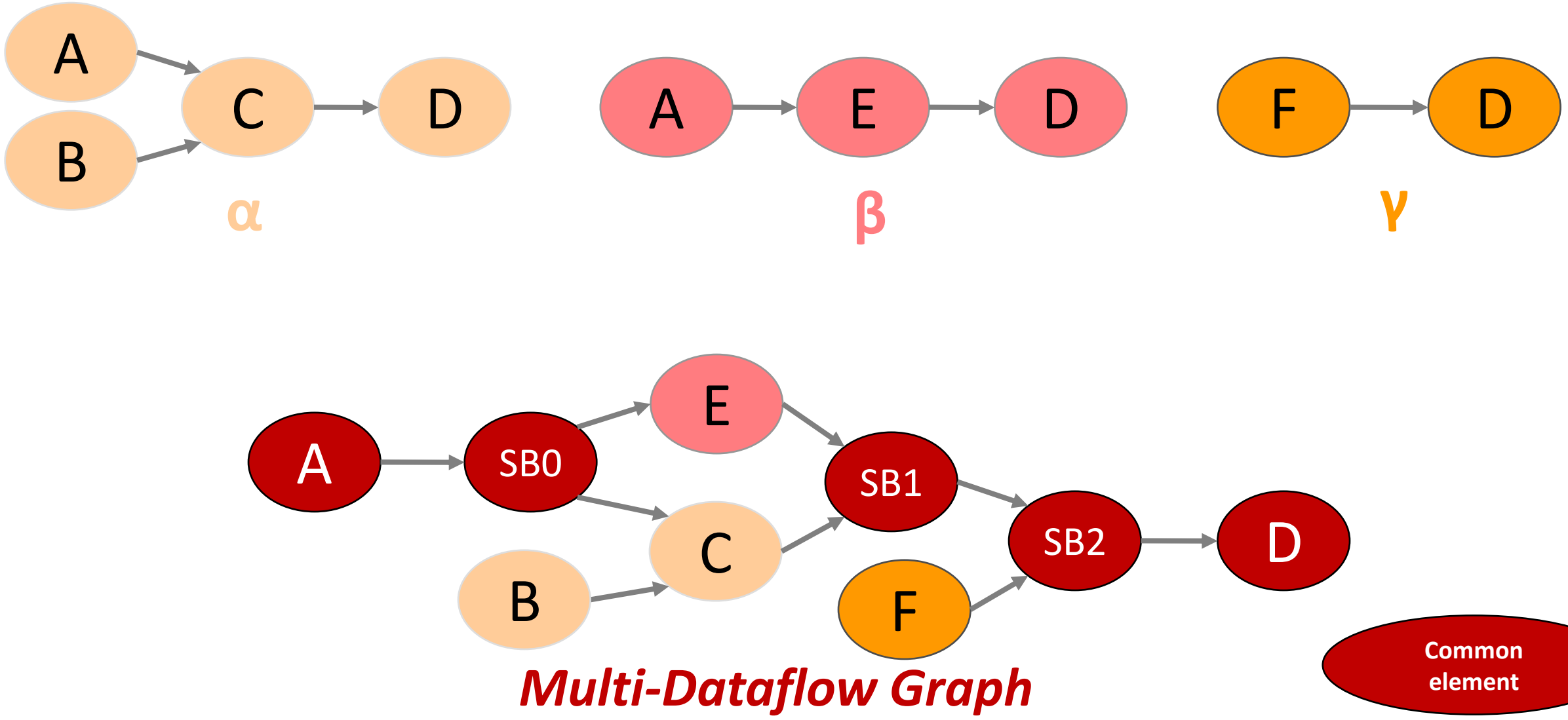
- *Technology part*: depends on the adopted technology. It defines the libraries to be used (for each operating conditions) and the cells (i.e. isolation, sleep transistors, state retentions) to be used within them.
- *Power intent part*: depends on the design. It defines all the rules to correctly operating the inserted switch (i.e. defines the on/off conditions of the sleep transistors), defines the set of available domains and which logic blocks belong to them

PSO Example



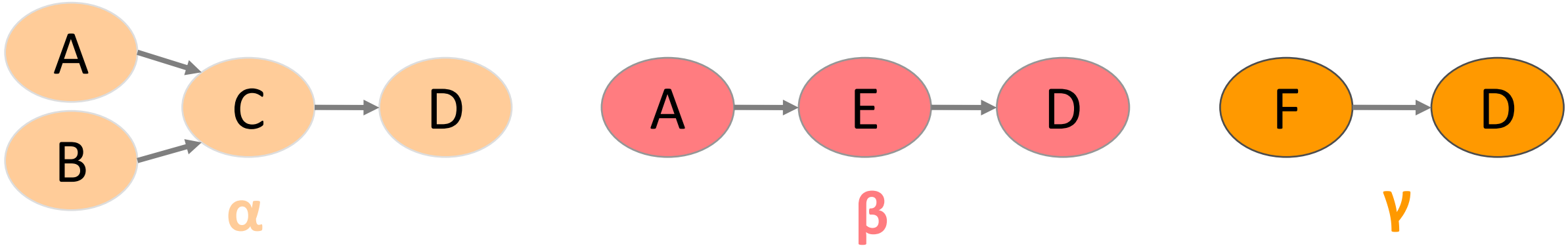
- Reconfigurable Filter, the depth of the filter can vary.
- 2 different Logic Regions (LR), one always on and one switchable
- Switchable LR needs the insertion of
  - isolation, retention and power switch cells;
  - one power controller to handle the control signals [1*shut-o + 1*isol. + 2*reten.]

Heterogeneous and Irregular CGR: power issue

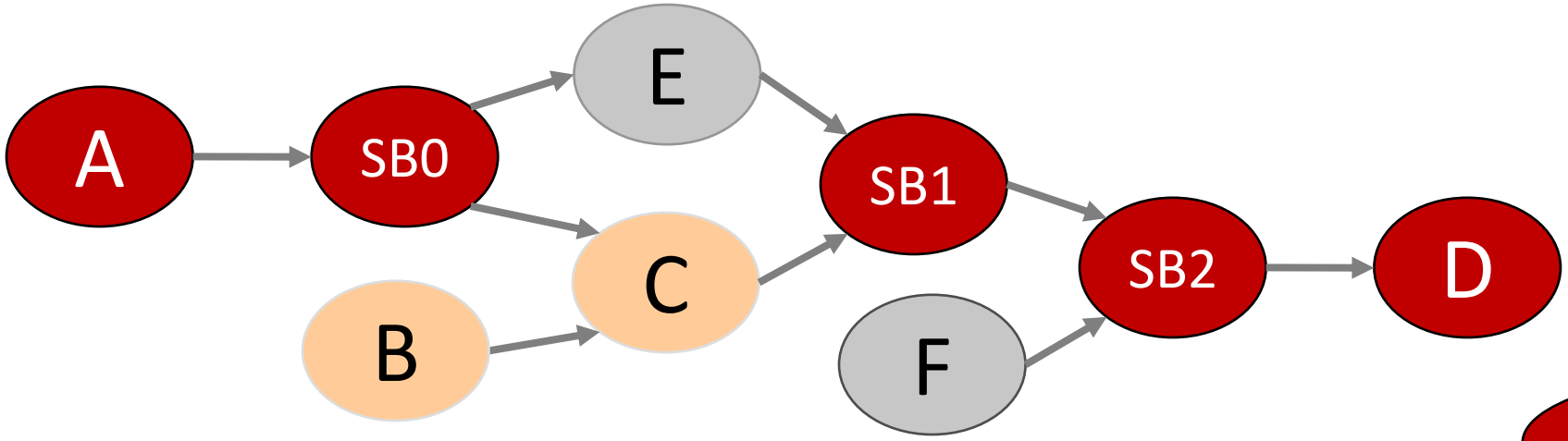


Multi-Dataflow Graph

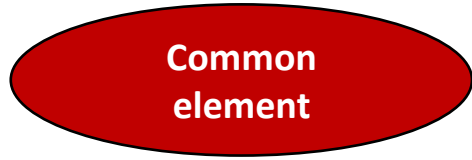
Heterogeneous and Irregular CGR: power issue



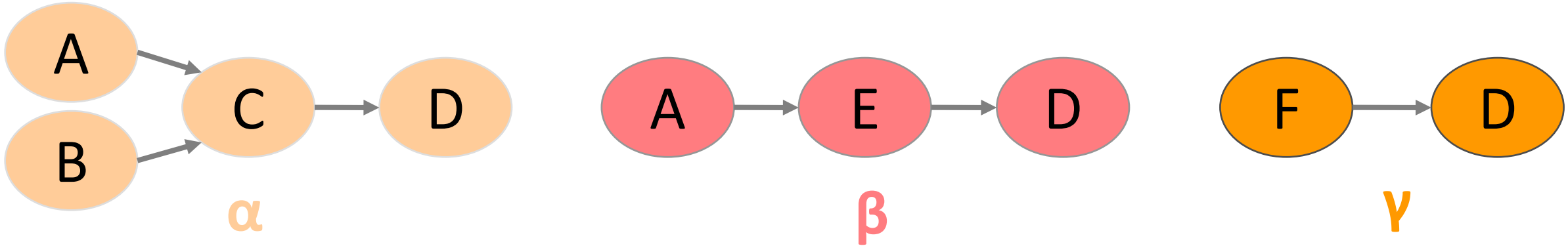
α execution: E and F, being not involved, are wasting power!



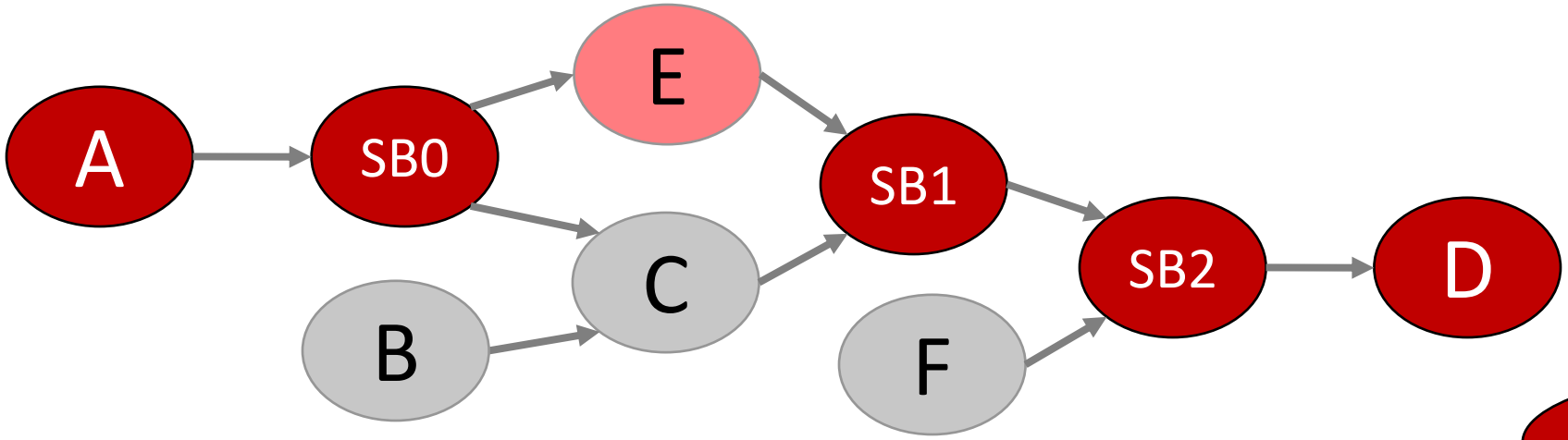
Multi-Dataflow Graph



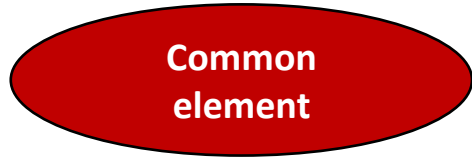
Heterogeneous and Irregular CGR: power issue



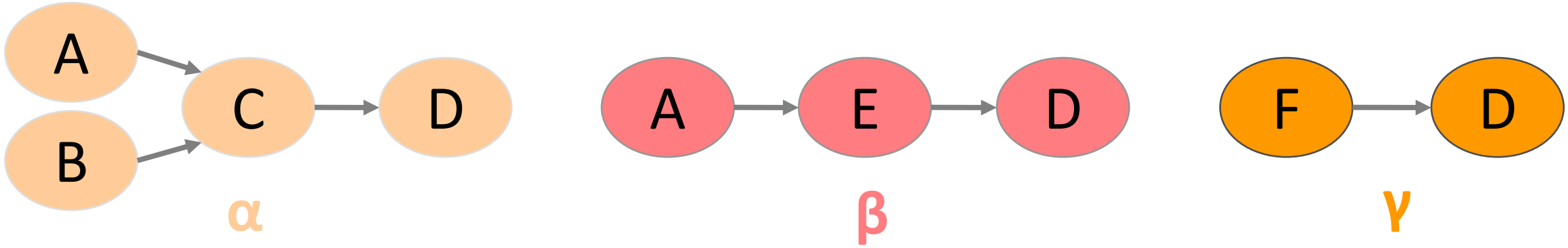
β execution: B C and D, being not involved, are wasting power!



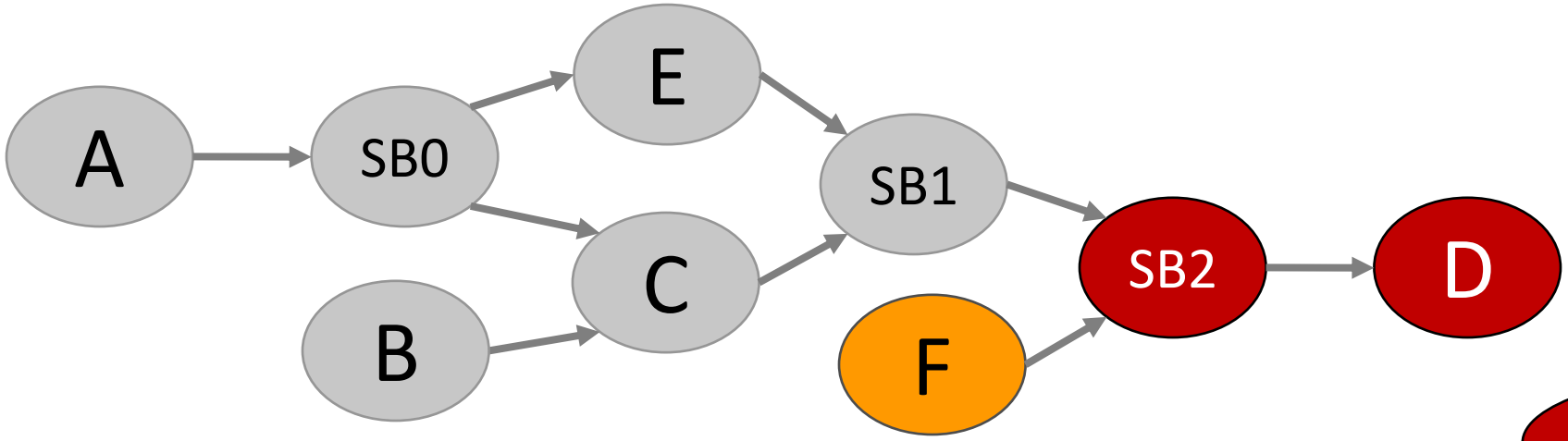
Multi-Dataflow Graph



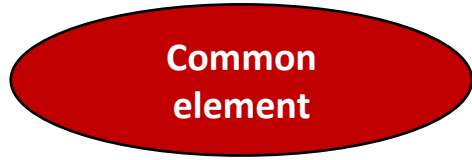
Heterogeneous and Irregular CGR: power issue



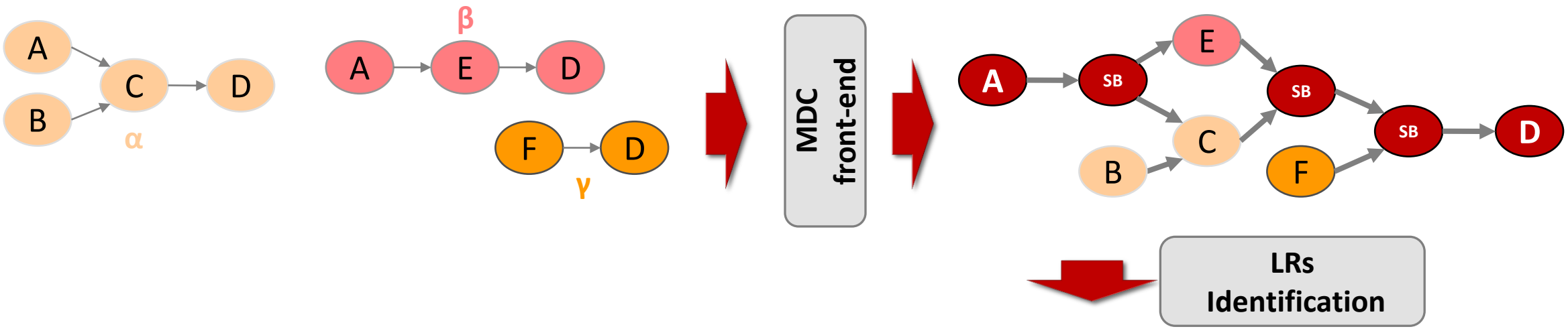
γ execution: B C and D, being not involved, are wasting power!



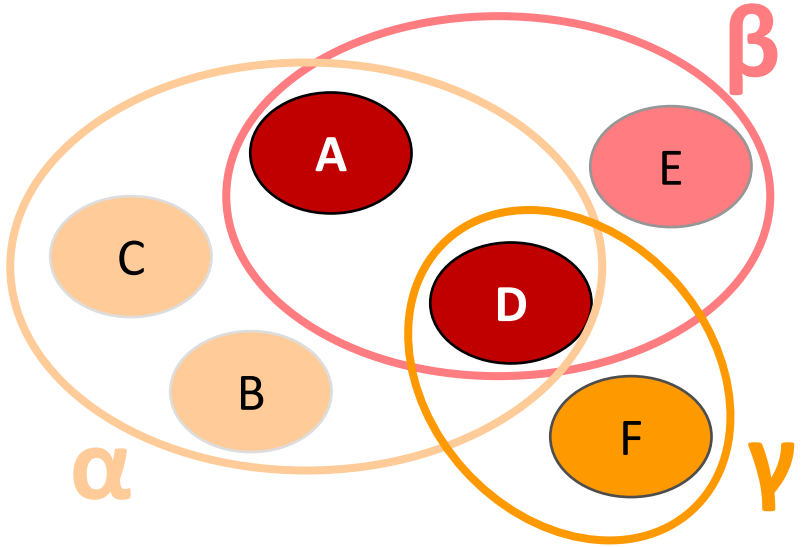
Multi-Dataflow Graph



Heterogeneous and Irregular CGR: MDC approach



LR	1	2	3	4	5
actors	A	B,C	D	E	F
α	1	1	1	0	0
β	1	0	1	1	0
γ	0	0	1	0	1

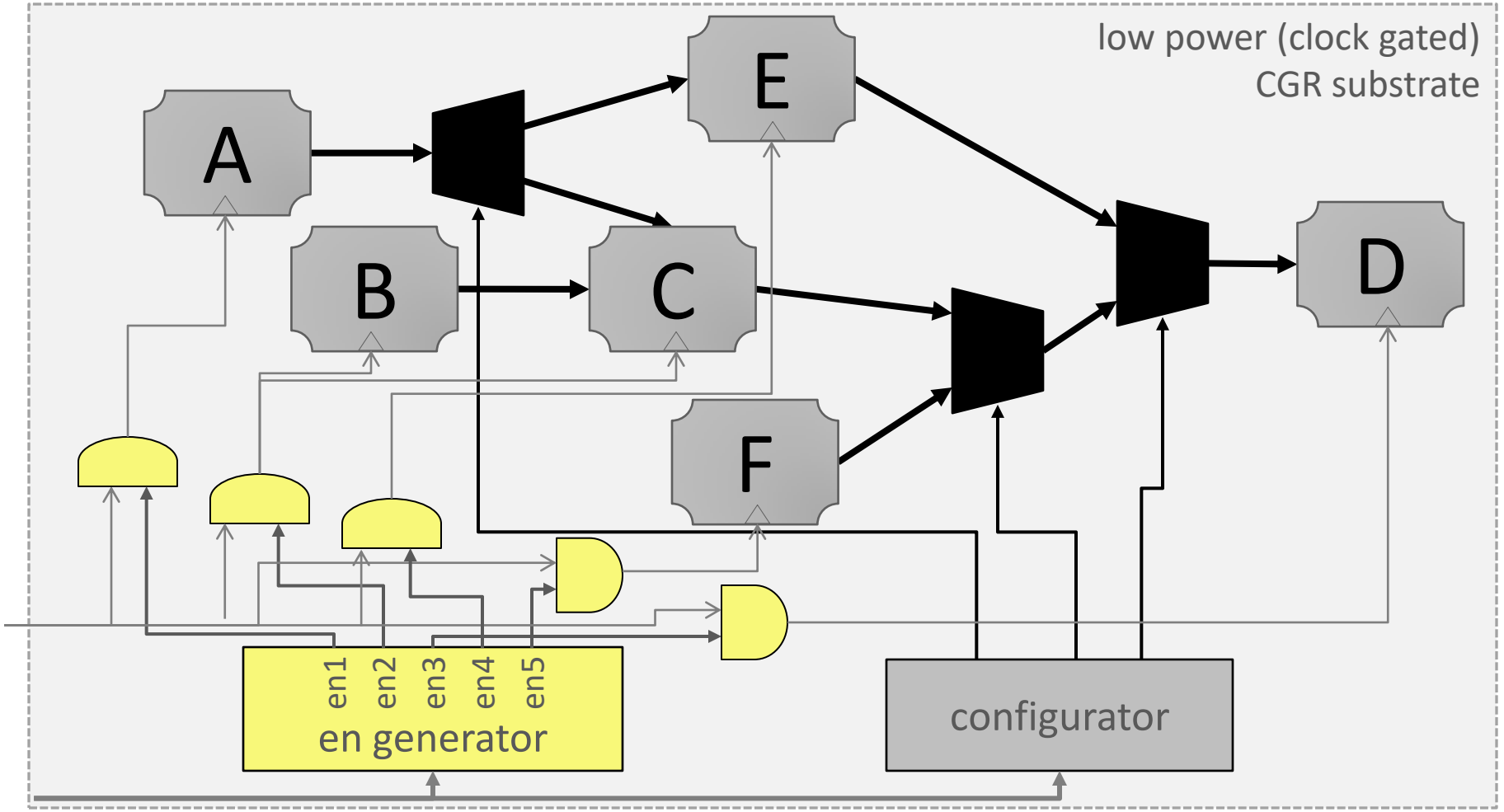


Heterogeneous and Irregular CGR: MDC approach

LR	actors	α	β	γ
1	A	1	1	0
2	B,C	1	0	0
3	D	1	1	1
4	E	0	1	0
5	F	1	0	1

MDC back-end

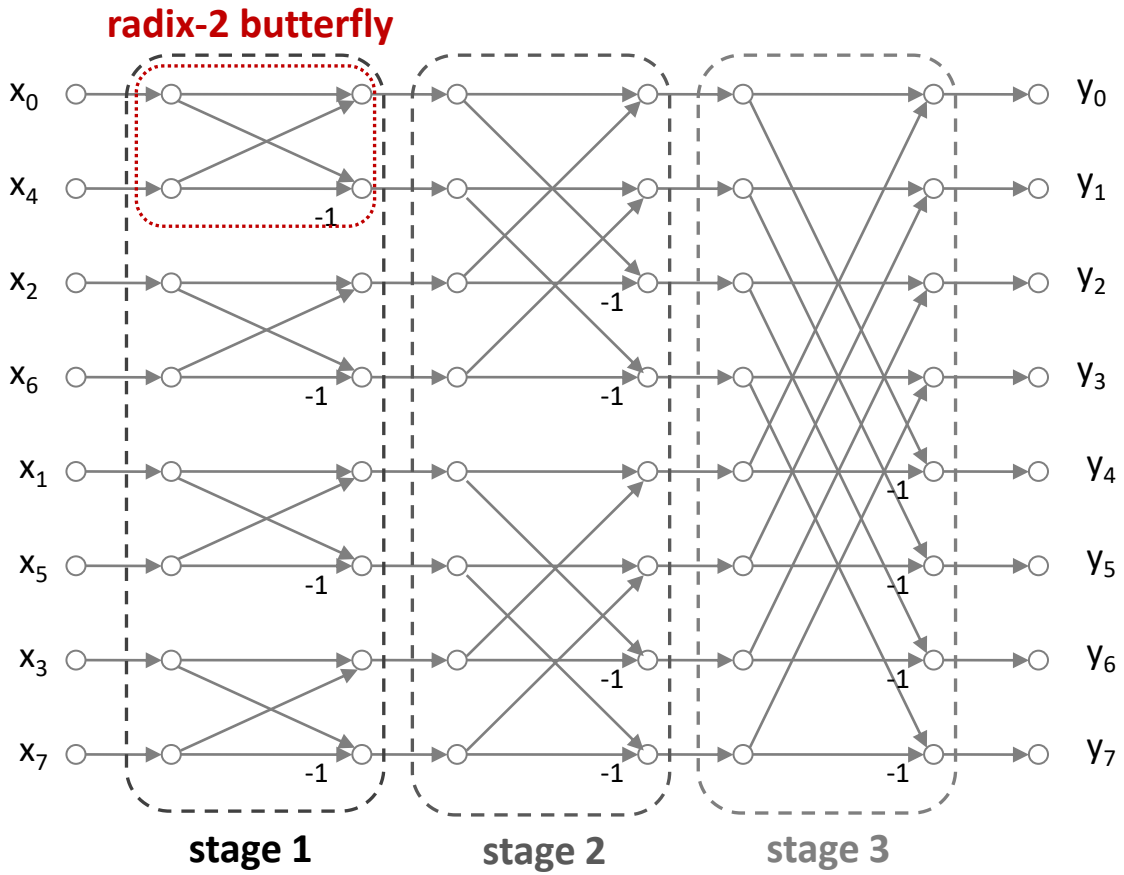
clk



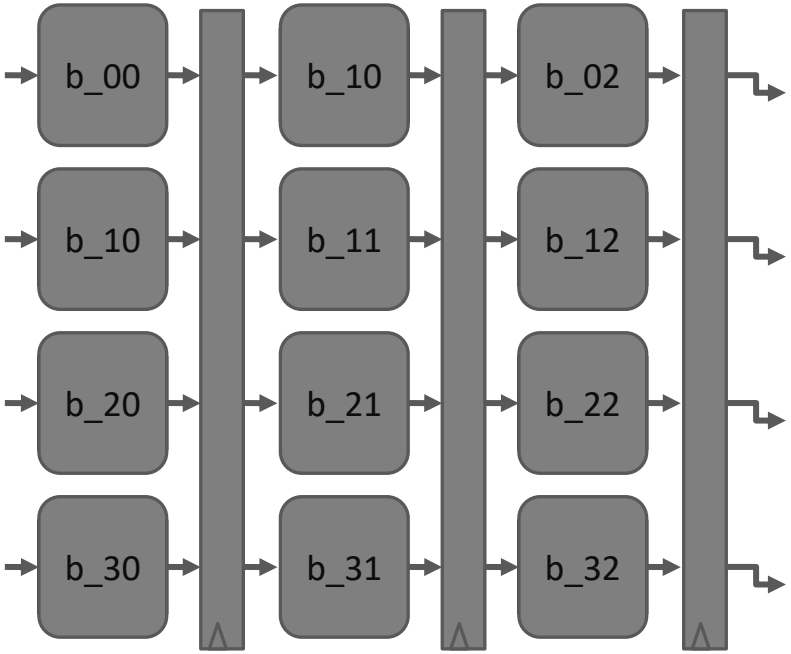
Overview

- Motivations
  - What we need adaptation for
  - Triggers and Types
- Coarse-Grained Reconfigurable Systems
  - Computing Spectrum and Reconfigurable Systems Classification
  - Heterogeneous and Irregular Coarse-Grained Reconfigurable Accelerators
- Power Management
  - Issues and Strategies
  - Low-Power Coarse-Grained Reconfigurable Accelerators
- Reconfigurable FFT Example and Conclusions

CG Reconfigurable FFT Design

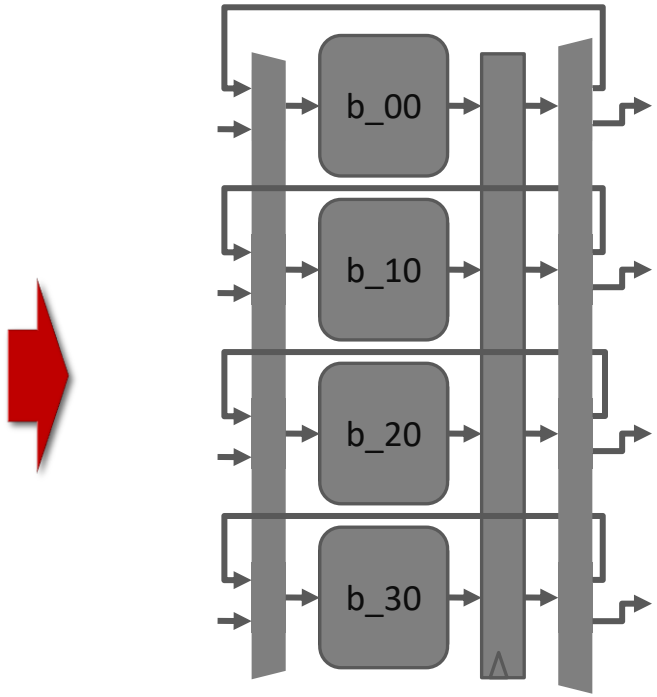


CG Reconfigurable FFT Design

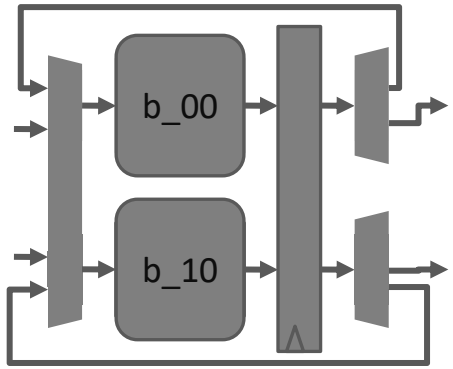


12 butterflies

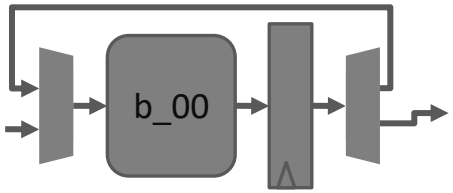
3cc



4 butterflies

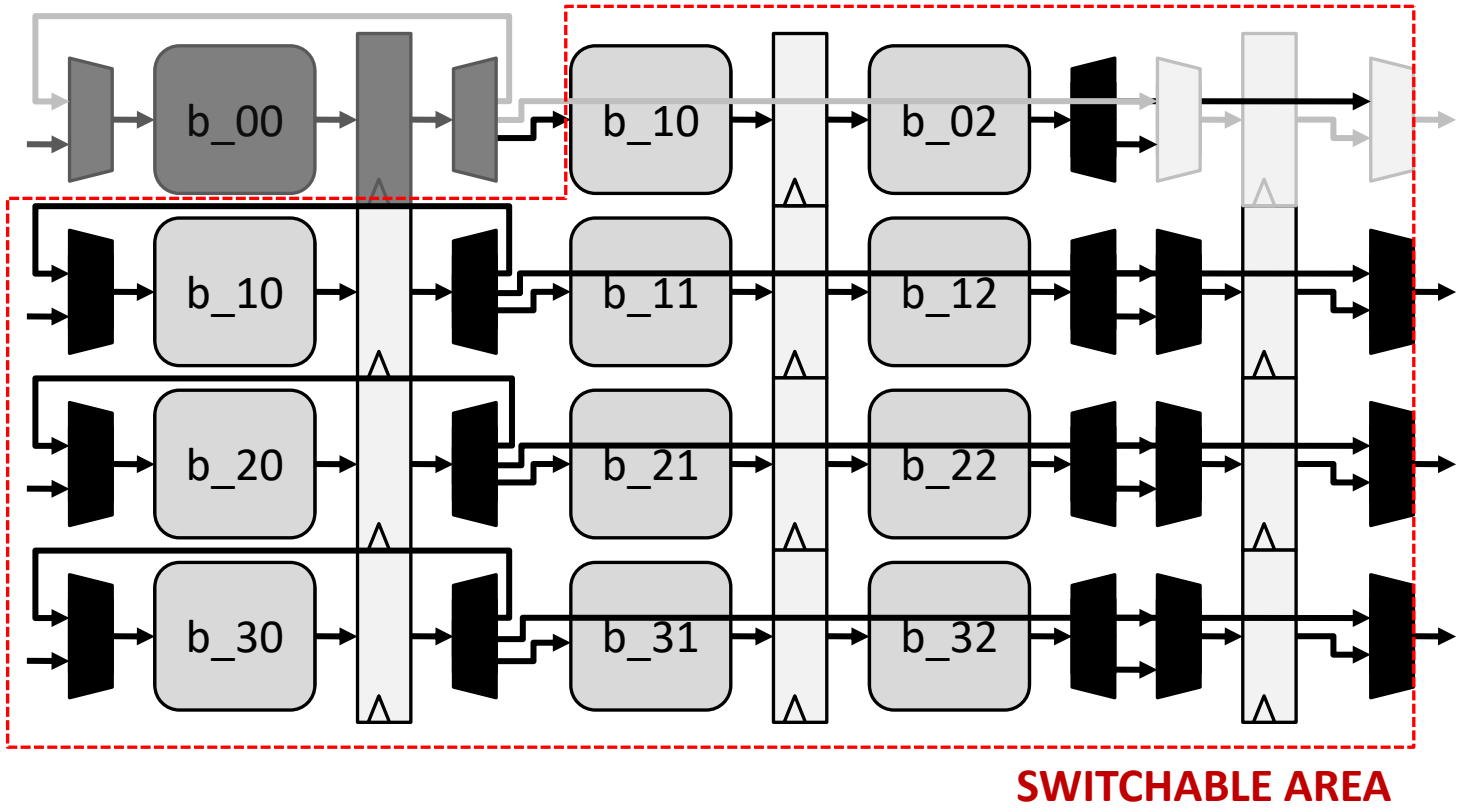


2 butterflies

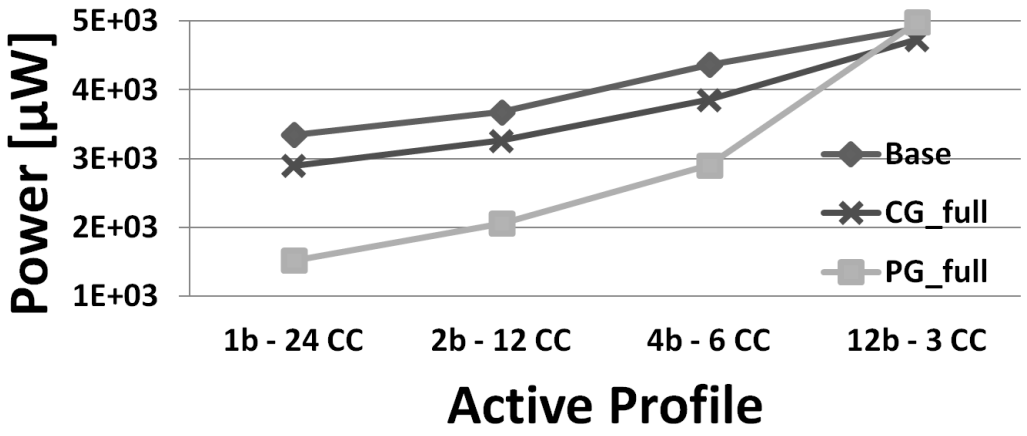


1 butterfly

Low-Power CG Reconfigurable FFT Design



Low-Power CG Reconfigurable FFT: 90nm ASIC



FFT: power vs throughput

Dynamic trade-off management

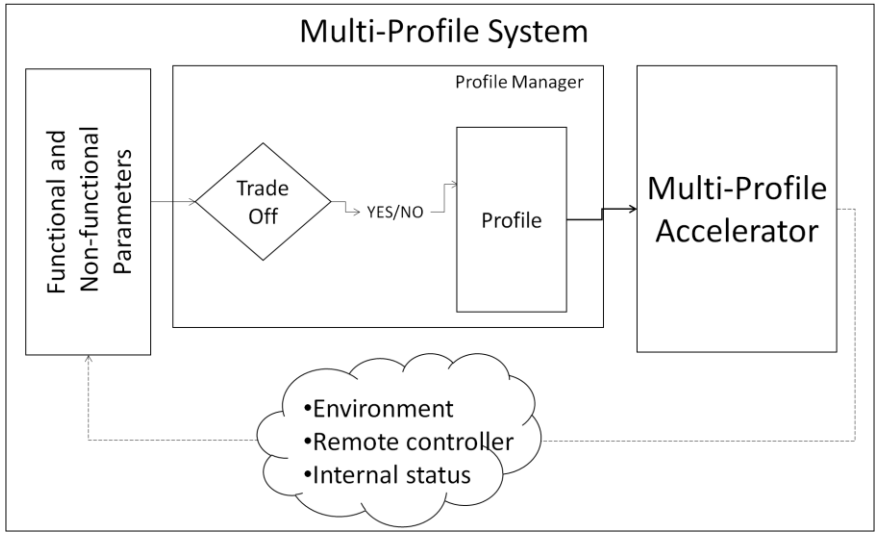
	Area [GE]	vs Base%
Base	885575	---
CG_full	885871	0.03
PG_full	926835	4.66

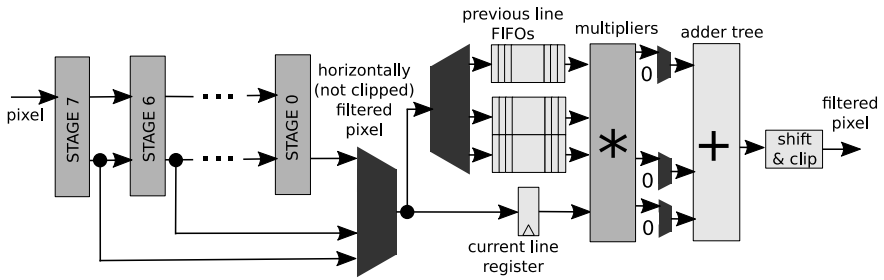
FFT: Area

MDC offers automatic implementation of power-gated and clock-gated designs

Final Remarks

- Reconfiguration is a good recipe to address flexibility
  - Functional: change the tasks is possible.
  - Non-Functional: change the execution profile is possible.
- Power is not necessarily an issue
  - The Multi-Dataflow composer tool offer ways to minimize consumption of unused resources, which a negligible impact/effort.





Example of multi-profile CGR system: HEVC interpolator

[ESL17] Carlo Sau, Francesca Palumbo, Maxime Pelcat, Julien Heulot, Erwan Nogues, Daniel Menard, Paolo Meloni, and Luigi Raffo. "***Challenging the Best HEVC Fractional Pixel FPGA Interpolators with Reconfigurable and Multi-frequency Approximate Computing***" in IEEE Embedded Systems Letters, vol. 9, no. 3, pp. 65-68, Sept. 2017.

Intelligent system DEsign and Application (IDEA) @ UNISS

Luca Pulina
Associate Professor

Francesca Palumbo
Assistant Professor

Arthur Bit-Monnot
Postdoctoral Researcher

Laura Pandolfo
Postdoctoral Researcher

Claudio Rubattu
Researcher Assistant

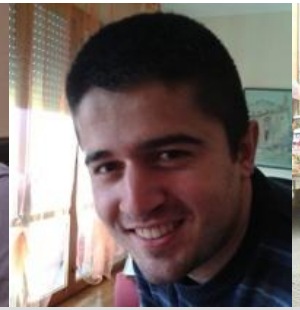
Simone Vuotto
Researcher Assistant

Monica Marini
Communication and Administration

The MDC Group – UNISS + UNICA team

UNIVERSITY OF SASSARI


UNIVERSITY OF CAGLIARI


CERBERO H2020 Project

EU Commission for funding the **CERBERO** (*Cross-layer modEl-based fRamework for multi-oBjective dEsign of Reconfigurable systems in unceRtain hybRid envirOnments*) project as part of the **H2020 Programme** under grant agreement No 732105.

Coordinator:

Michal Masin (IBM), michaelm@il.ibm.com

Scientific Coordinator:

Francesca Palumbo (UniSS), fpalumbo@uniss.it

Innovation Manager:

Katiuscia Zedda (Abinsula), katiuscia.zedda@abinsula.com

Dissemination-Communication Manager:

Francesco Regazzoni (USI), francesco.regazzoni@usi.ch

www.cerbero-h2020.eu

info@cerbero-h2020.eu

@CERBERO_h2020

Design for Low-Power IoT Systems:
CG Reconfigurable Acceleration Units

FRANCESCA PALUMBO

UNIVERSITÀ DEGLI STUDI DI SASSARI