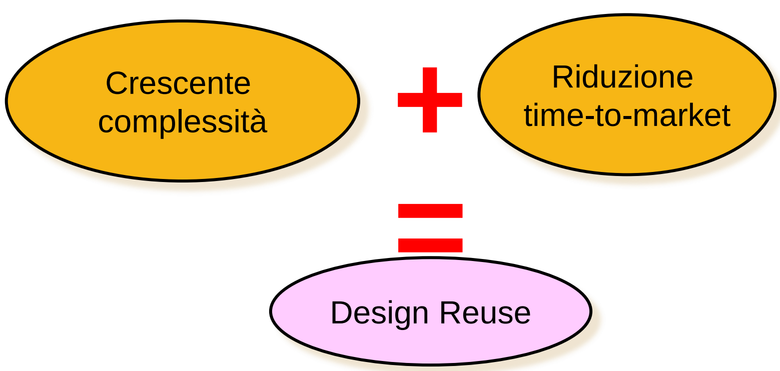
Università di Catania
Dipartimento di Ingegneria Elettrica Elettronica ed Informatica

Design Space Exploration: a parameterized VLIW platform

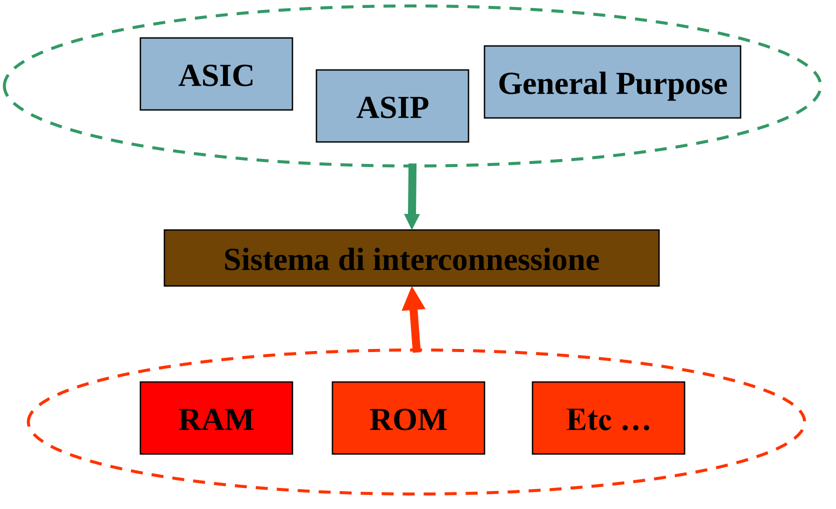
Ing. Davide Patti

Trend nella progettazione



- **Riconfigurazione di blocchi preesistenti (IP cores)**
- **Platform-based design**

Processing + Storage elements

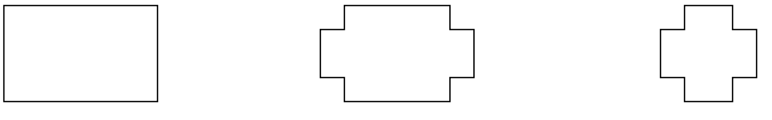


Processing Elements

- Processors vary in their customization for the problem at hand

Desired functionality

```
total = 0
for i = 1 to N  loop
   total += M[i]
end loop
```



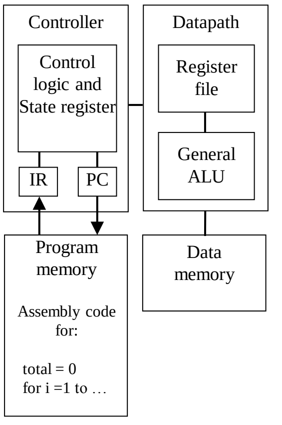
General-purpose processor

Application-specific processor

Single-purpose processor

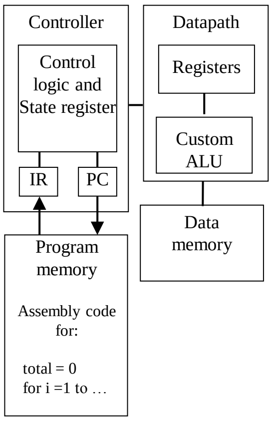
General-purpose processors

- Programmable device used in a variety of applications
  - Also known as "microprocessor"
- Features
  - Program memory
  - General datapath with large register file and general ALU
- User benefits
  - Low time-to-market and NRE costs
  - High flexibility
- Intel/AMD the most well-known, but there are hundreds of others



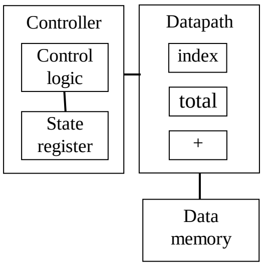
Application-specific processors

- Programmable processor optimized for a particular class of applications having common characteristics
  - Compromise between general-purpose and single-purpose processors
- Features
  - Program memory
  - Optimized datapath
  - Special functional units
- Benefits
  - Some flexibility, good performance, size and power

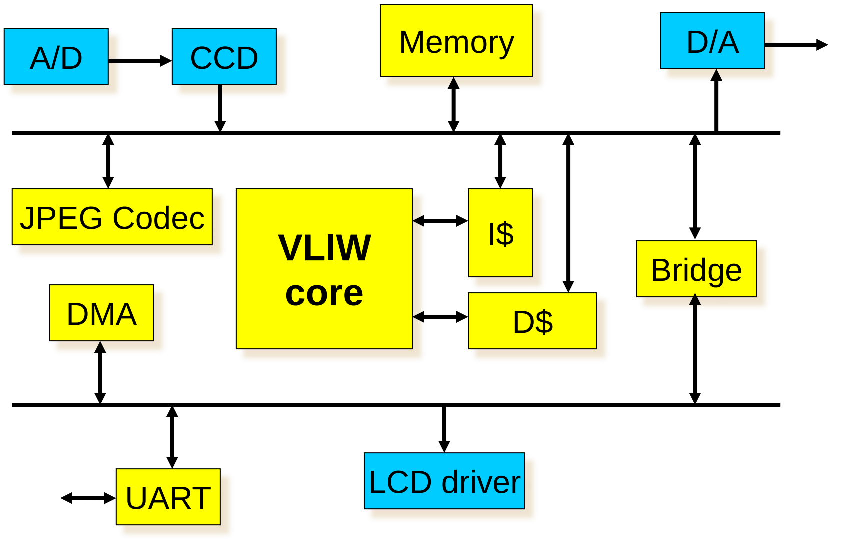


Single-purpose processors

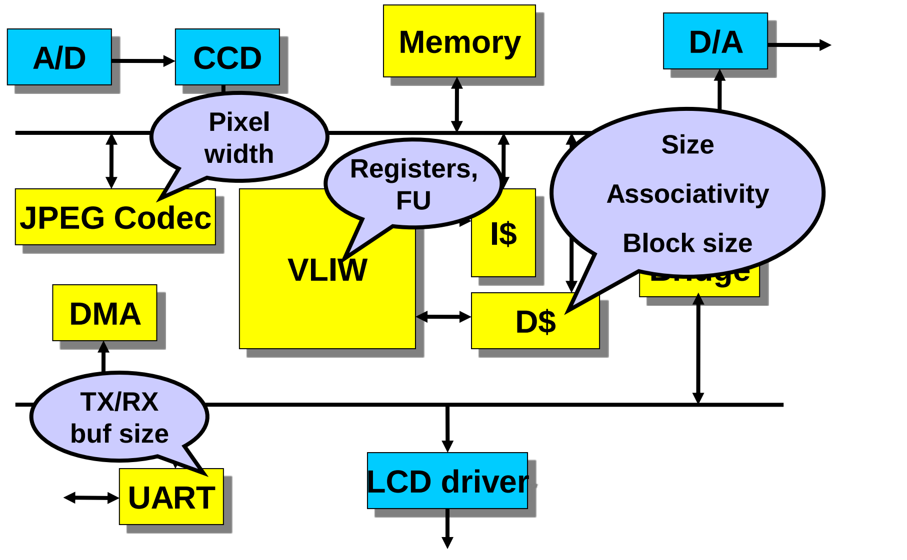
- **ASIC (application specific integrated circuit)**: Digital circuit designed to execute exactly one program
  - a.k.a. coprocessor, accelerator or peripheral
- Features
  - Contains only the components needed to execute a single program
  - No program memory
- Benefits
  - Fast
  - Low power
  - Small size



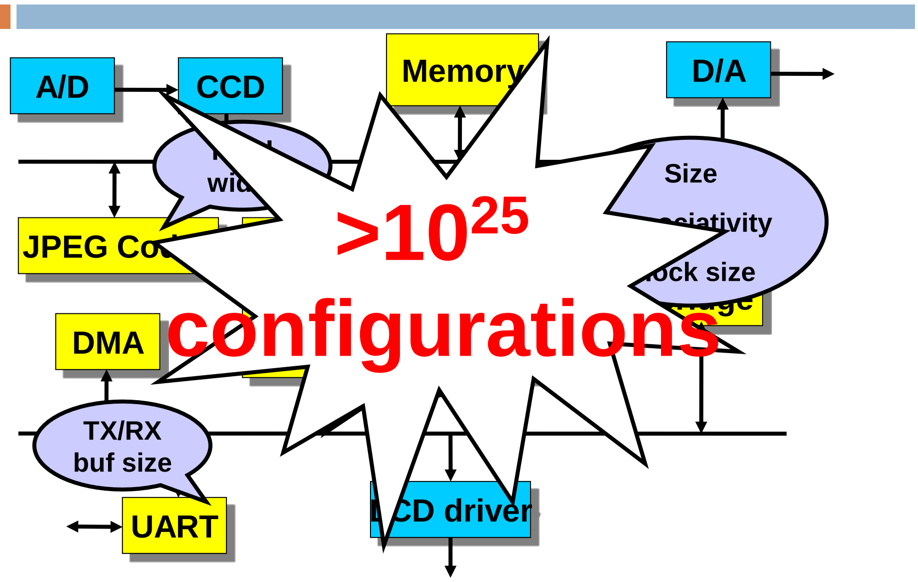
Digital Camera Example



Sample SOC Platform for Digital Camera



Sample SOC Platform for Digital Camera

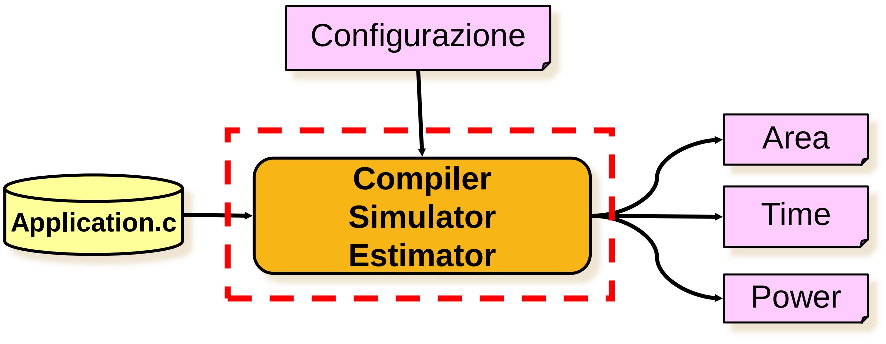


Parameterized Platforms

- For such architectures to be reused for various applications they have to be heavily parameterized
  - ➔ Parameterized computational, communication, and memory elements
- Terminology
  - ➔ A complete assigment of values to all the parameters is a **configuration**
  - ➔ A complete collection of all possible configurations is the Configuration Space (a.k.a., the Design Space)

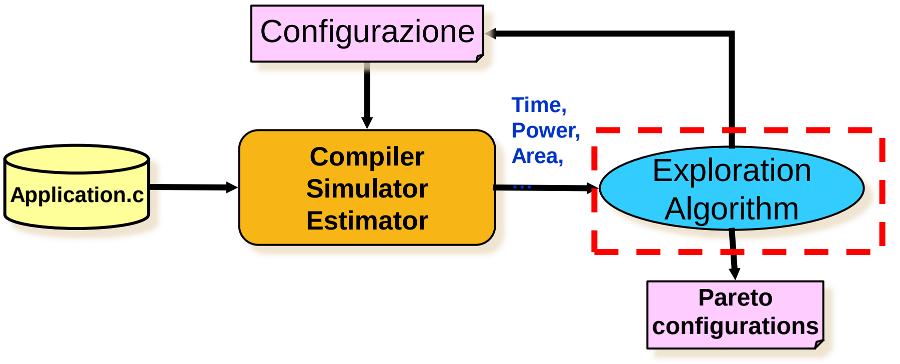
Strumenti necessari (1/2)

- Implementazione di modelli di stima ad alto livello per una rapida valutazione delle grandezze obiettivo

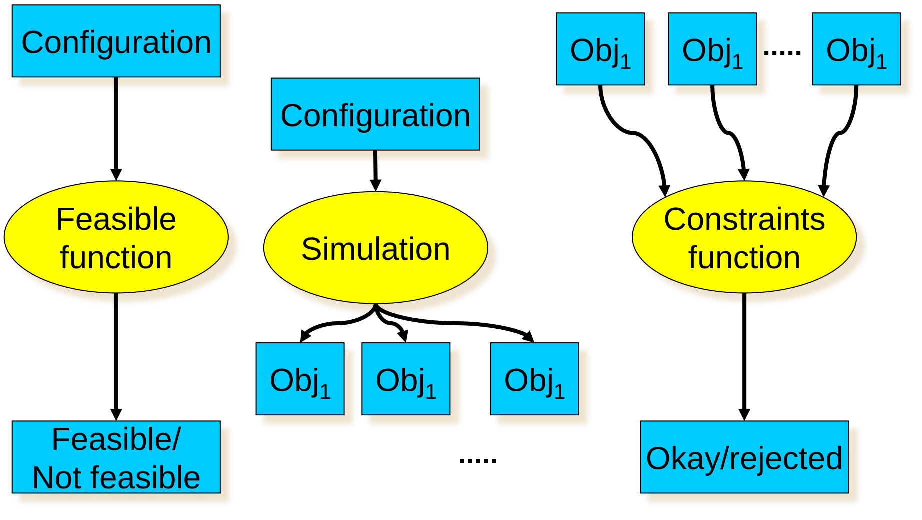


Strumenti necessari (2/2)

- **Una strategia di esplorazione intelligente dello spazio delle configurazioni**



Feasible/Constraints Functions

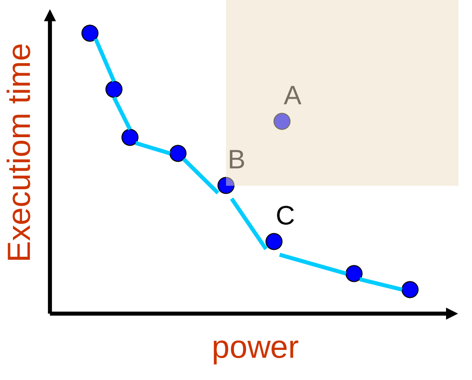


Design Space Exploration (DSE)

- Defining strategies for tuning the parameters so as to obtain the Pareto-optimal set of configurations that provide multi-criteria optimisation
- Criteria (a.k.a. objectives)
  - Power dissipation
  - Performance (delay, execution time, …)
  - Area (cost, complexity)
  - Energy
  - …

Pareto's Concept

- A new notion of optimality is required in the presence of objective conflicts

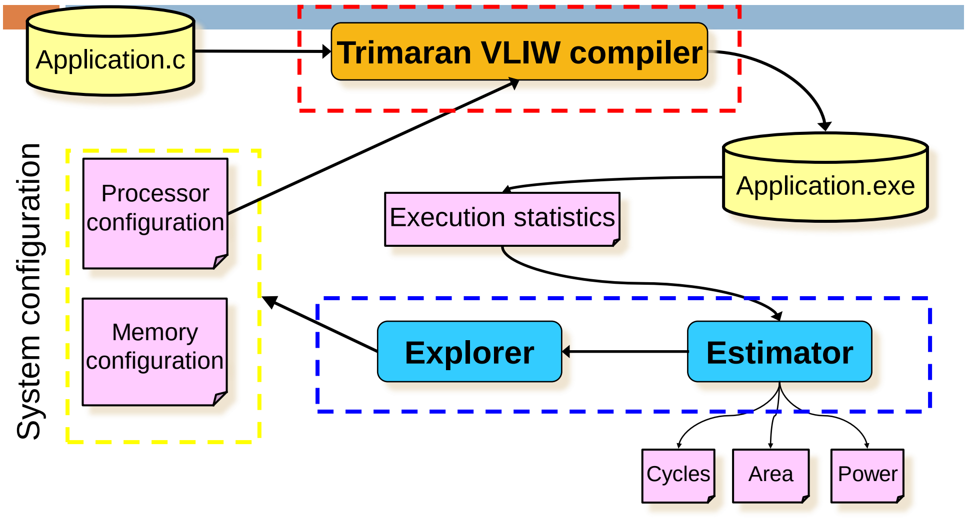


Piattaforma EPIC Explorer

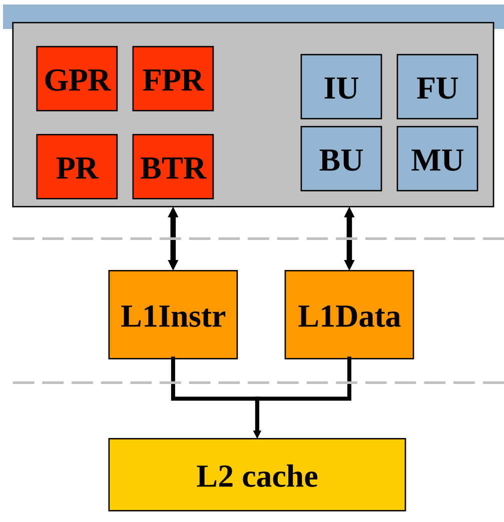
- Interfacciamento al framework di compilazione VLIW **Trimaran** (HP Labs, ReaCT-ILP Laboratory at NY University)
- Integrazione **modelli di stima** performance/power/area
- Sviluppo **algoritmi di esplorazione** dello spazio di progetto
- **Open platform**: sviluppata su GNU/Linux e liberamente disponibile con licenza GPL

`code.google.com/p/epic-explorer/`

EPIC Explorer: Flusso dei dati



Architettura di riferimento

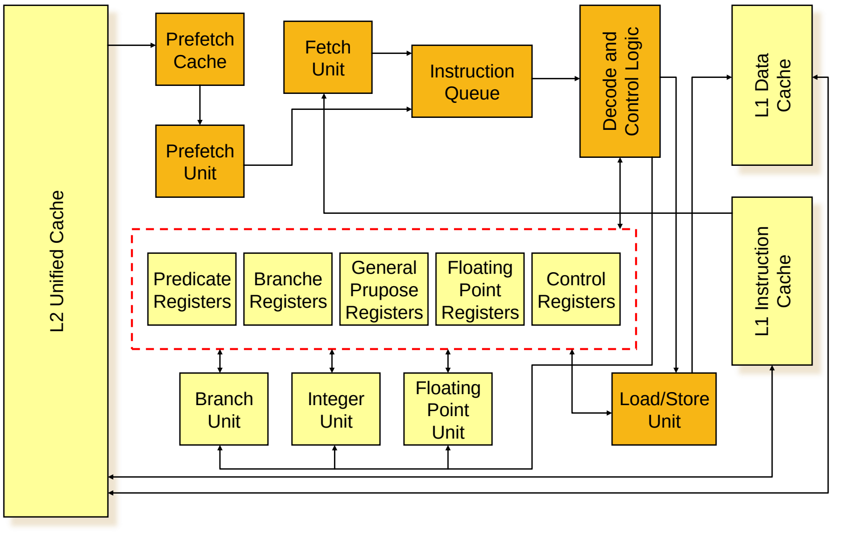


EPIC/VLIW core
- Unità funzionali
- Register files

Cache di Livello 1

Cache di Livello 2

Reference architecture (HPL-PD)



Energy Estimation

- Processor (Functional Units, Register Files)
- Caches
- Buses

Goals of the used models :

✓ Discrete degree of accuracy (about 25%)

✓ Demonstrate relative power savings beetween designs

Energy estimation

- Subdivide architecture in *Functional Block Unit* (FBU)
  - → Instruction decode logic, Integer units, floating point units, register files etc..
- For each FBU (from ST Microelectronics LX)
  - → **Active power:** average power dissipated when the FBU is used
  - → **Inactive power:** average power dissipated when the FBU is not used (usually ranges from 10 to 50% of active power)
- From the execution statistic, we know how many cycles each FBU has been active/inactive
  - → $E_{FBU}=(P_{active}\times cycles_{active}+P_{inactive}\times cycles_{inactive})\times T_{clock}$
- Discrete degree of accuracy (about 25%)
  - → investigate relative power savings beetween designs

Power Estimation (buses)

- Bus lines transitions computed from the list of data/address memory accesses

$$P_{bus} = 0.5 \times (V_{dd})^2 \times \alpha \times f \times C_l$$

- → V_{dd} supply voltage
- → α switching activity
- → f clock frequency
- → C_l capacity of a bus line

Attività

- Ottimizzazione multi-obiettivo
  - → Modelli di stima power/performance/area
  - → Estrazione Pareto-Set
- Analisi & Sviluppo di algoritmi per il Design Space Exploration
  - → Efficienza temporale
  - → Accuratezza dei risultati
- Valutazione della politica di compilazione
  - → Stima degli effetti sugli obiettivi di progetto

Reference Application Set

- MediaBench suite

Application	Category
`G721 encode`	Voice compression
`Gsm encode`	Speech transcoding
`Gsm decode`	Speech transcoding
`Ieee 810`	IEEE 1180 inverse DCT
`JPEG`	Image compression
`MPEG2 decode`	Video decoding
`ADPCM encode`	Speech encoding
`ADPCM decode`	Speech decoding
`Fir`	FIR filter

Configuration Space

Three main parameter categories:

- **VLIW core:**
  - → Number of Registers in each register file (from 16 to 256)
  - → Number of istancies for Functional Units of each type (from 1 to 6)
- **Mem Hierarchy:**
  - → Size, Blocksize, Associativity for each of the caches (L1 Instruction, L1 Data, L2)
- **Compiler:**
  - → Conservative compilation strategy (basic blocks)
  - → Aggressive ILP oriented compilation strategy (hyperblocks)

Total space size: **1.47×10^{13}** configurations!

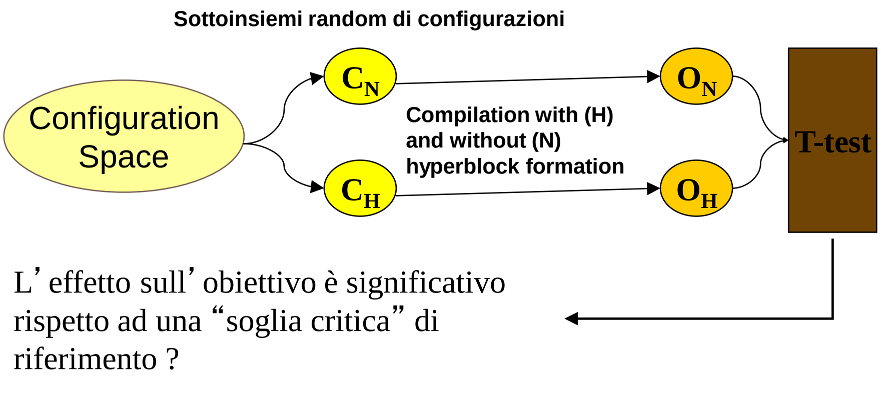
Metodologia di esplorazione

- Analisi preliminare della compilazione
  - Impatto sugli obiettivi di trasformazioni ILP-oriented
  - Eventuale scelta di una politica di compilazione
    - Basic Blocks (conservative)
    - Hyper Blocks (aggressive, ILP-oriented)

- Multi-objective Design Space Exploration
  - Analisi & Clustering dei Pareto-Set

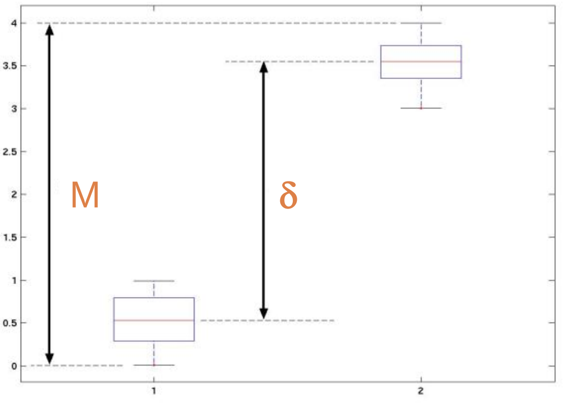
Analisi preliminare

- Per ciascun obiettivo, si è utilizzato un ***Unpaired two sample t-test*** per stimare l'effetto medio di una compilazione ILP oriented.



Analisi preliminare

- Example of a metric for critical difference in means: **δ > 50% M**

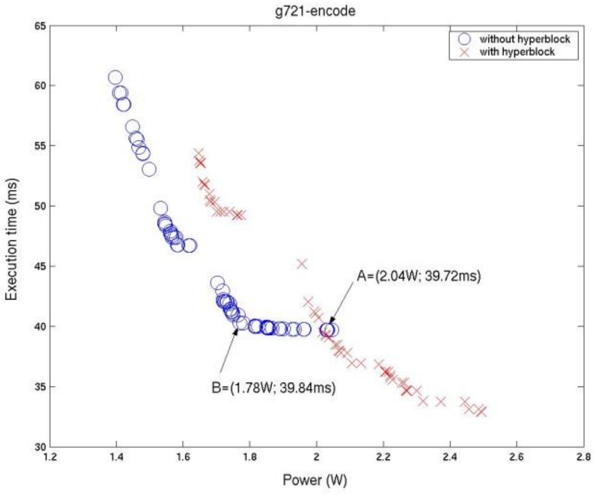


Stima impatto sugli obiettivi

ILP-oriented compilation impact (positive,negative)

Application	Time (ms)		Power (W)		Energy (mJ)	
	Δ	μ_N-μ_H	Δ	μ_N-μ_H	Δ	μ_N-μ_H
ieee810	16.64	6.76±1.84	1.64	0.38 ± 0.16	**49.01**	**30.82** ± **4.55**
gsm-enc	**36.62**	**33.25** ± **4.79**	0.88	-0.48 ± 0.14	**79.28**	**55.84** ± **9.82**
jpeg	4.07	-0.97 ± 0.51	0.89	-0.07 ± 0.09	9.72	-2.31 ± 1.01
adpcm-enc	15.8	8.17 ± 2.2	1.25	-0.89 ± 0.14	46.12	-8.56 ± 3.73
MPEG dec	33.39	-5.28 ± 4.85	0.88	0.25 ± 0.16	62.50	-3.48 ± 9.88
G721-enc	22.76	-7.23 ± 2.95	0.76	-0.39 ± 0.08	65.53	-32.4 ± +5.9
adpcm-dec	24.2	-6.19 ± 3.31	1.02	-0.5 ± 0.12	58.54	-27.74 ± 7.3
Fir	0.68	-0.26 ± 0.08	0.79	-0.27 ± 0.09	1.40	-0.97 ± 0.12
gsm-dec	21.55	-23.83 ± 2.58	0.54	-0.24 ± 0.09	59.60	-56.6 ± 6.43

Pareto Set (G721 encode)



Design Space Exploration

Even using fast high level estimation models, we need "intelligent" exploration strategies to avoid exhaustive evaluation of all possible configurations.

Two main goals of DSE:

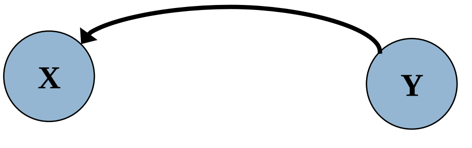
- **Accuracy:** results similar to exhaustive exploration.
- **Efficiency:** optimal pareto set searched in a reasonable time.

Exploration Algorithms

- Dependency analysis (dep), Givargis *et al.*, [TVLSI'02]
- GA-based DSE (ga), Palesi *et al.*, [CODES'01]
- Sensitivity Analysis, Fornaciari *et al.*, [DAES'02]
  - Pareto-based Sensitivity Analysis (pbsa), Palesi *et al.*, [VLSI-SOC'01]

Dependency analisys

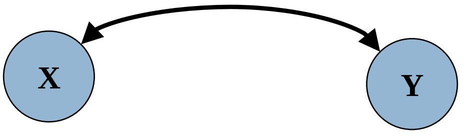
- If the optimal value of a parameter X depends on the value of an other parameter Y, the X is said dependent from Y.



- Optimal values of X must be computed once optimal values of Y have been computed

Dependency analisys

- If X depends on Y, and Y depends on X, parameters are defined interdependent.

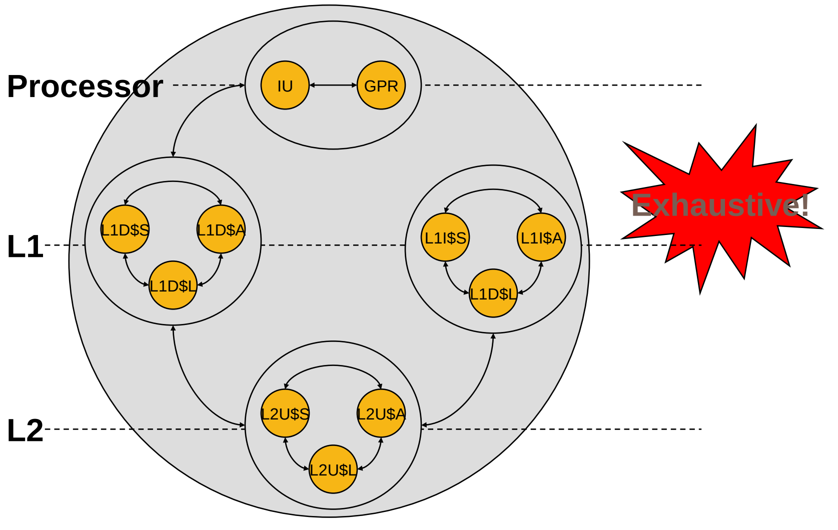


- The optimal values of interdependent parameters must be computed simultaneously.

dep: How It Works

- Interdependent parameters are grouped in clusters
- 1st phase
  - ➔ Clusters are exhaustively explored with the aim to compute the local Pareto-optimal set (LPOS)
- 2nd phase
  - ➔ The LPOSs are merged and exhaustively searched to find the global Pareto-optimal set (GPOS)

dep: Dependency graph



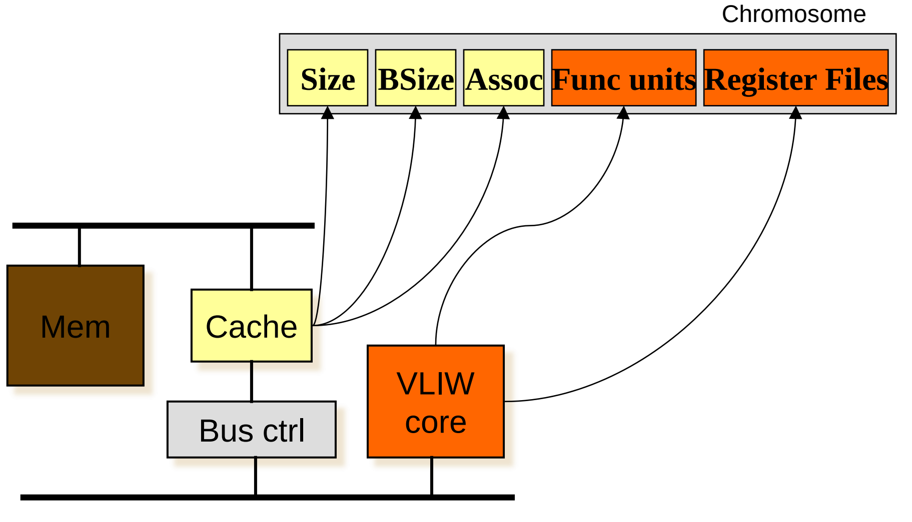
Sensitivity Analysis

- Minimization power-delay (*PD*) of a cache
  - A configuration is a triple $c=<s,b,a>$ (size,bsize,assoc)
  - Sensitivity analysis
    - Fix *b* and *a* and let *s* variable ⇒ PD^s_{min}, PD^s_{max}
    - Fix *s* and *a* and let *b* variable ⇒ PD^b_{min}, PD^b_{max}
    - Fix *s* and *b* and let *a* variable ⇒ PD^a_{min}, PD^a_{max}
    - **Sorting** starting <u>from the most sensitive to the less one</u> (e.g., *s*, *a*, *b*)
  - Exploration
    - Fix $a=a_0$, $b=b_0$ and make *s* variable ⇒ s_{opt}
    - Fix $s=s_{opt}$, $b=b_0$ and make *a* variable ⇒ a_{opt}
    - Fix $s=s_{opt}$, $a=a_{opt}$ and make *b* variable ⇒ b_{opt}

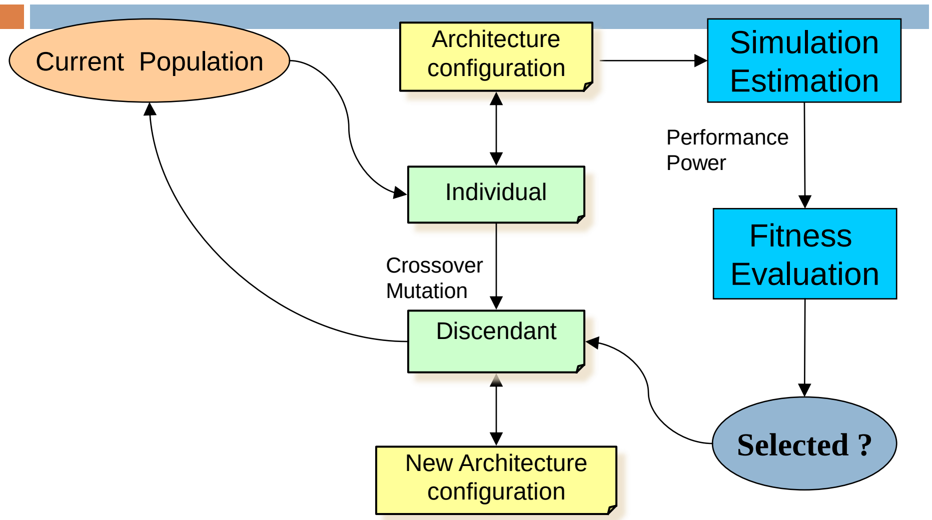
ga: Genetic Approach

- 5 items
  - Configuration representation
  - Feasible function
  - Cost/Objective functions
  - Constraint functions
  - Convergency criteria

ga: Configuration Representation



DSE: Genetic Iteration

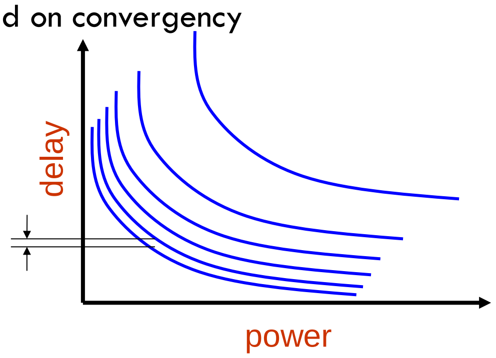


Multiobjective Fitness assignment

- Strength Pareto Approach [Zitzler,Thiele]
- From current population P is extracted an ***external set P**** containing the *nondominated* configuration of P.
- **Fitness of P* element j:** $f_j = n/(N+1)$
  - ➔ N = total size of P
  - ➔ n = # of P configurations dominated by j
- **Fitness of P element i:** 1/S .
  - ➔ S is the sum of the fitness values of the P* elements that dominates i

How Many Generations?

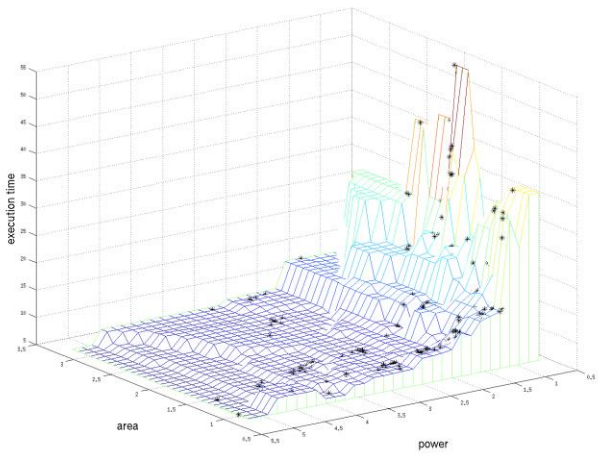
- Fixed number of generations
- Autostop criteria
  - ➔ Based on convergency



Vantages of Genetic Approach

- Dependency analisys is not required
- Customizable: population size, crossover probability, mutation probability etc.)
- Good efficiency: exploration time does not explode with larger parameters ranges
- Good accuracy:in the subspaces where it was possible to compare it to ehxaustive exploration it showed very good accuracy even with 5-10 generations.
- Generality of the approach: large number works already present in literature

Pareto Surface



DSE: sviluppi futuri

- Aggiornamento/creazione modelli di stima System-level
  - ➔ Sintesi e valutazione in VHDL
- Valutazione impatto delle tecniche di compilazione sugli obiettivi
- Algoritmi di esplorazione