



MULTICUBE Explorer: Leveraging DoE/RSM-based Techniques to Automate Design Space Exploration for CMPs

Cristina Silvano

Politecnico di Milano, Milano (ITALY)
Dipartimento di Elettronica e Informazione
silvano@elet.polimi.it
<http://home.dei.polimi.it/silvano/>

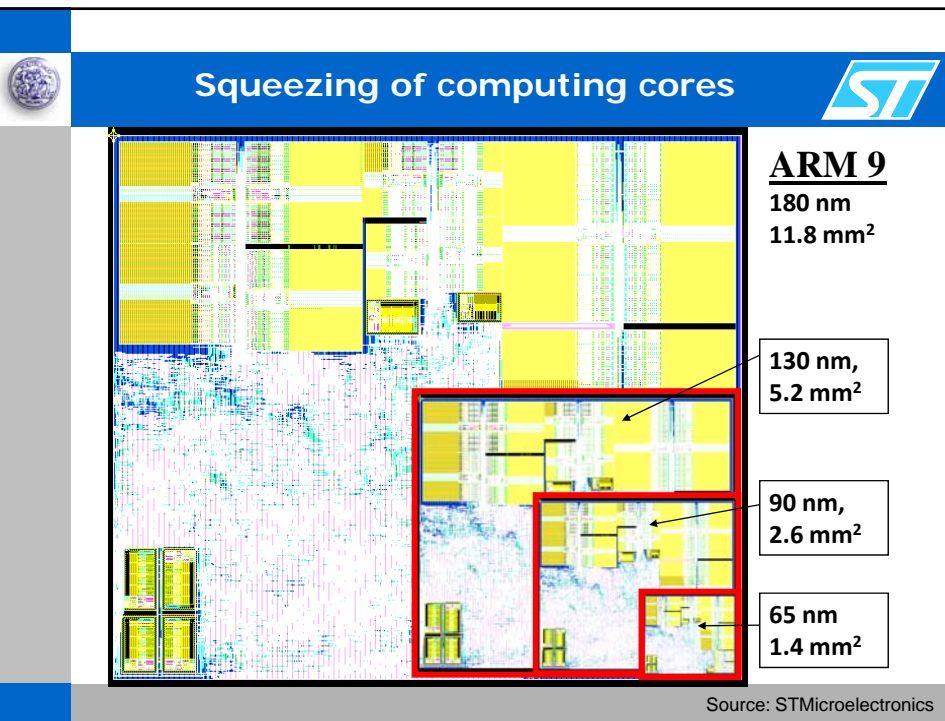


Outline

- Introduction and Motivations
- Automatic Design Space Exploration Methodology
 - Design of Experiments
 - Response Surface Modeling
- Experimental Results
- MULTICUBE Explorer Open-source Tool
- Conclusions



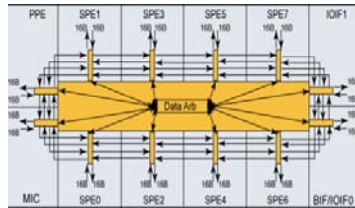
Introduction and Motivations



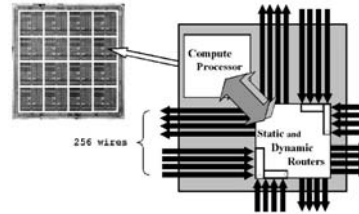


From multi-core to many-core architectures

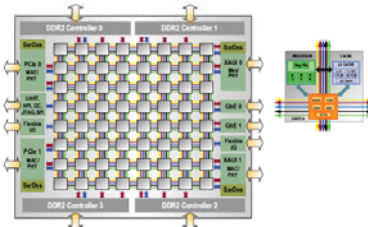
IBM - Cell/B.E.



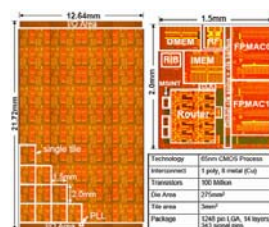
MIT - RAW



Tilera - TILE64



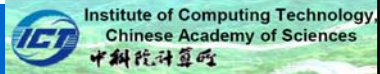
Intel - Terascale



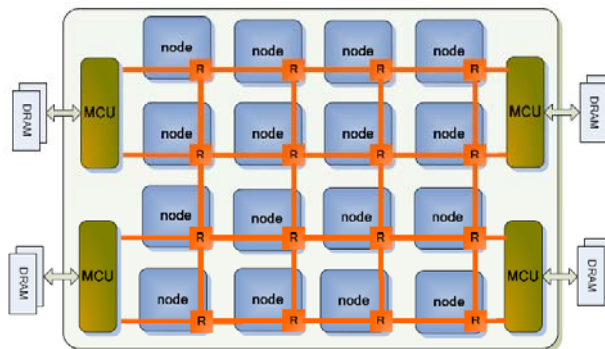
Cristina SILVANO - Politecnico di Milano (ITALY)



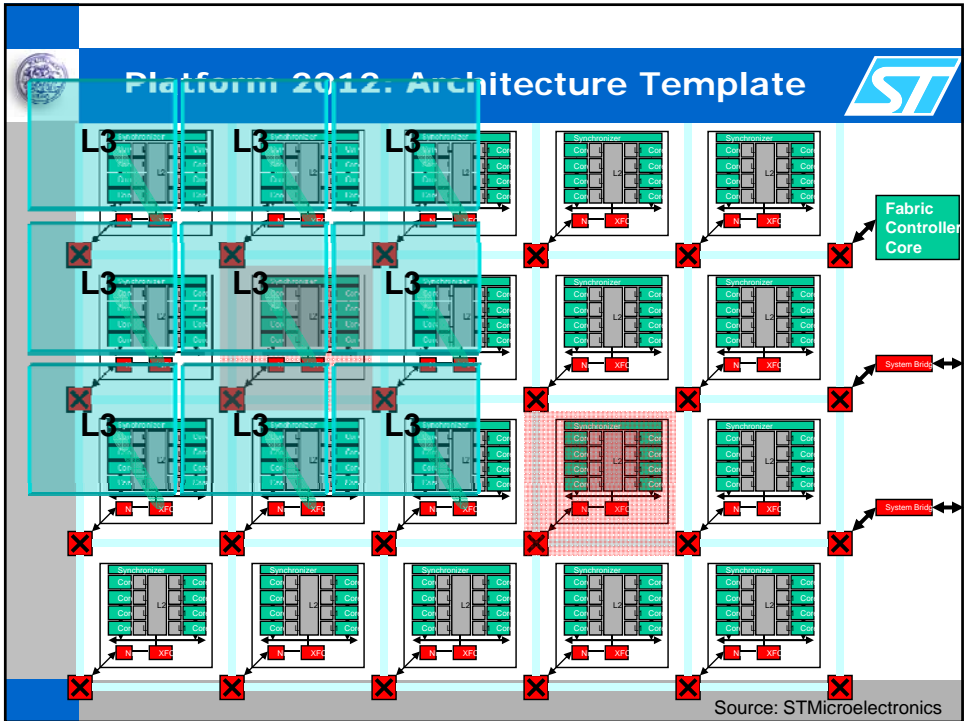
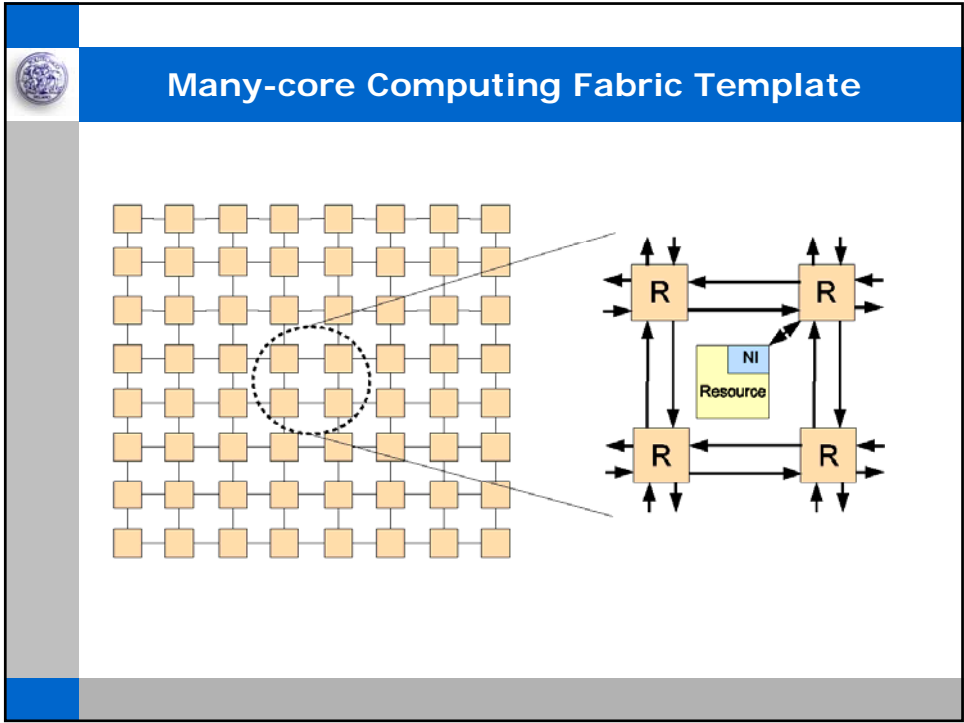
Multi-core architecture



- Multi-core architecture: a tiled homogeneous multi-core architecture for general embedded purpose (**Godson-T**)



Source: ICT (ICT is partner of MULTICUBE project)





Introduction and Motivation

- Given the increasing **complexity** of Chip Multi-Processors, a wide range of architecture parameters (number of processors, processor issue width, L1 & L2 cache size, etc.) must be tuned
- Design space of the target architecture A should consider all possible configurations of each parameters p_i :

$$A = S_{p1} \times S_{p2} \times \dots \times S_{pn}$$

- Example:
 - Number of Processors = {2, 4, 8, 16}
 - Processor Issue Width = {1, 2, 4, 8}
 - L1 Instr. Cache Size = {2KB, 4KB, 8KB, 16KB}
 - L1 Data Cache Size = {2KB, 4KB, 8KB, 16KB}
 - L2 Private Cache Size = {32KB, 64KB, 128KB, 256KB}
 - L1 Instr. Cache Associativity = {1-way, 2-way, 4-way, 8-way}
 - L1 Data Cache Associativity = {1-way, 2-way, 4-way, 8-way}
 - L2 Data Cache Associativity = {1-way, 2-way, 4-way, 8-way}
 - I/D/L2 Cache Block Size = {16, 32}

⇒ Huge design space composed of 2^{17} (131 072) system configurations

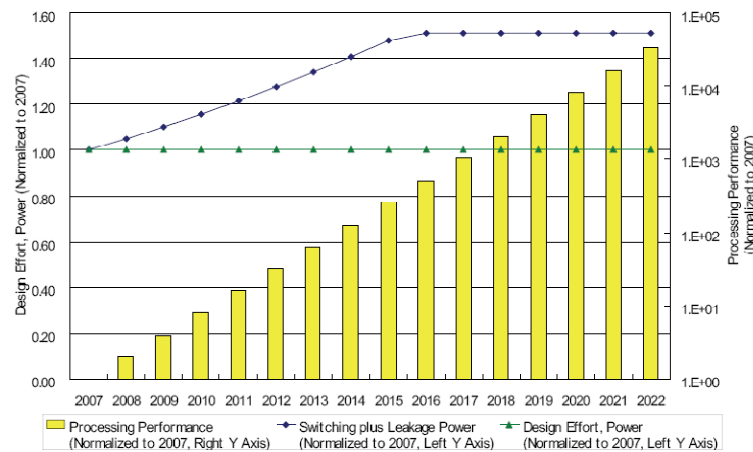
9

Cristina SILVANO - Politecnico di Milano (ITALY)



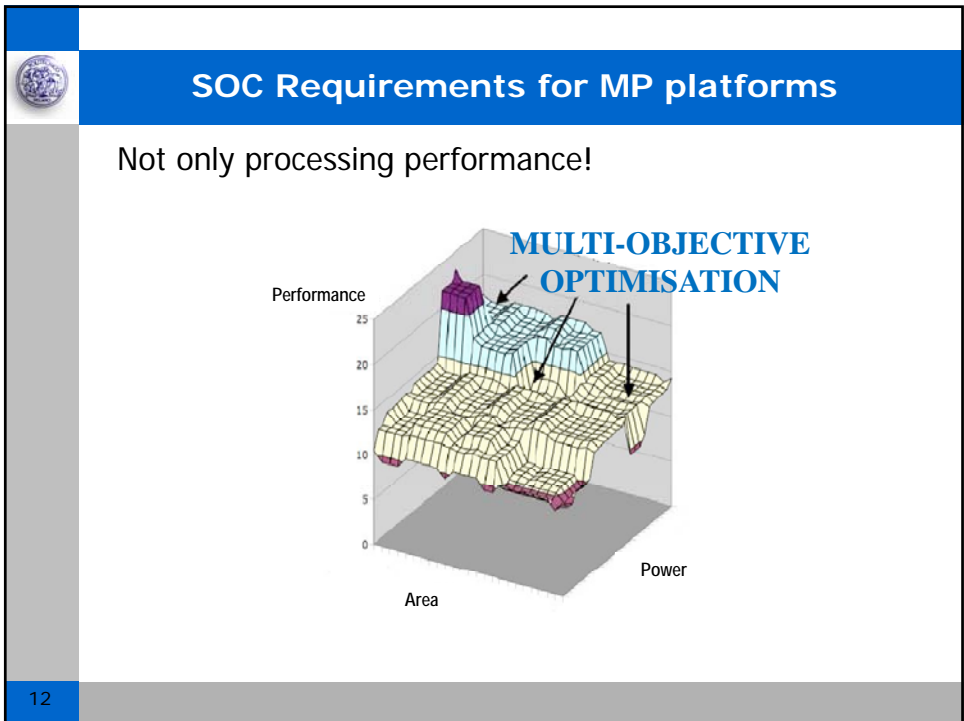
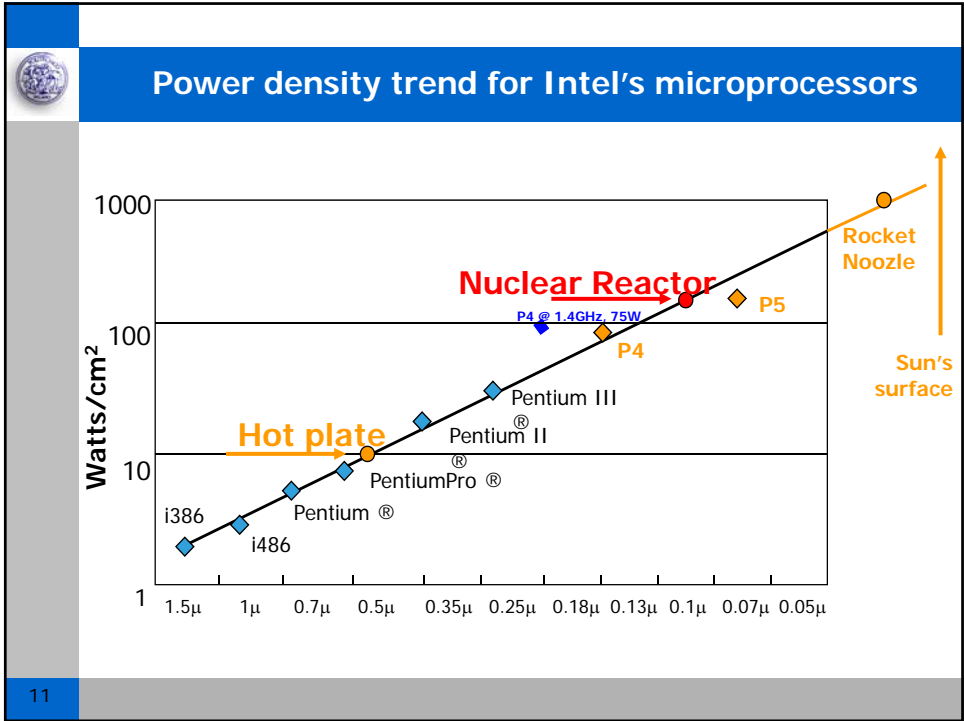
SOC Requirements for MP platforms

Processing performance is expected to grow more than 2 orders of magnitude in the next 10 years.



10

Source: STMicroelectronics





Introduction and Motivation

- Given the increasing complexity of Chip Multi-Processors, a wide range of **architecture parameters** (number of processors, issue width, L1 & L2 cache size, etc.) must be explored to find the best trade-off in terms of **multiple objectives** (energy, delay, bandwidth, area, etc.).
- **Multi-Objective Exploration** of the huge design space of next generation CMPs cannot be anymore based on intuition and past experience of the design architects
- **Need for Automatic Design Space Exploration** to support systematically the exploration and the quantitative comparison in terms of multiple competing objectives

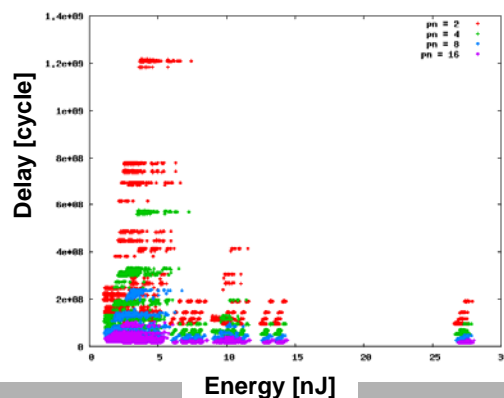
13

Cristina SILVANO - Politecnico di Milano (ITALY)



Full Search Design Space Exploration

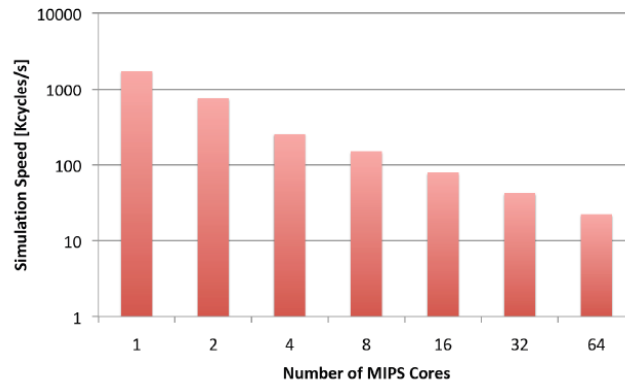
- In most cases, the design space to be explored is **huge**
- Automatic Design Space Exploration based on **full-search exploration** is unfeasible because it requires a very long simulation time
- Example: design space composed of 131 072 system configurations.
If simulation of the target application for each system configuration requires 1 min. \Rightarrow 131 072 min \sim 91 days for full-search exploration



14



SESC simulation speed by varying the number of MIPS cores



What is SESC?

SESC is an open-source cycle accurate architectural simulator of MIPS instructions set. It models single processors and several configurations of CMPs. SESC project started at University of Illinois at Urbana-Champaign





MULTICUBE Project

- So far, several heuristic techniques have been proposed to address the design space exploration problem, but they are all characterized by low efficiency
- An overall design space exploration framework is needed to combine all optimizations into a global search space with a common interface to the simulation and evaluation tools.
- **MULTICUBE FP7-ICT Project** focuses on the definition of an **automatic multi-objective Design Space Exploration (DSE) framework** to be used to tune Chip Multi-Processor architectures evaluating a set of metrics (such as energy and delay) for the next generation embedded computing platforms.



Automatic Design Space Exploration










 

MULTICUBE Project

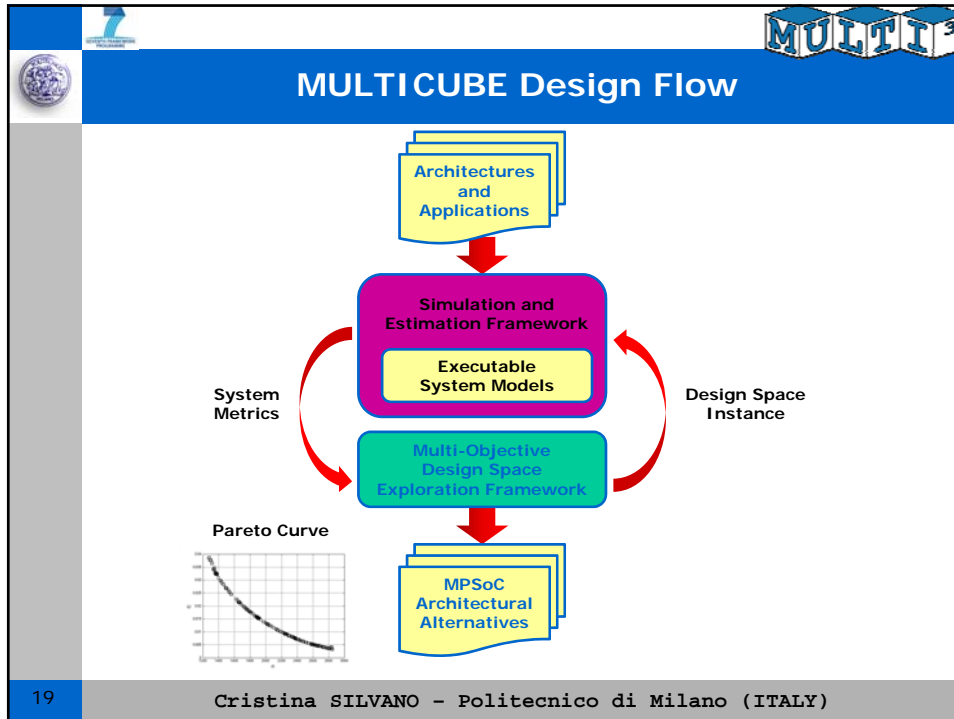
MULTI-OBJECTIVE DESIGN SPACE EXPLORATION OF MULTI-PROCESSOR SOC ARCHITECTURES FOR EMBEDDED MULTIMEDIA APPLICATIONS

www.multicube.eu

Project Duration: from January 2008 to June 2010

 Politecnico di Milano (POLIMI) - Italy (Project Coordinator)	 Università della Svizzera Italiana (ALaRI) - CH
 DS2 - Spain	 University of Cantabria - Spain
 STMicroelectronics - Italy	 STMicroelectronics - China
 imec - Belgium	 Institute of Computing Technology (ICT) - China
 ESTECO - Italy	

18 **Cristina SILVANO - Politecnico di Milano (ITALY)**



-
- The slide lists the main goals of the work. It begins with the introduction of a **Multi-Objective Design Space Exploration (DSE) framework** for customizing MP-SoC architectures. The framework is simulation-based and focuses on **Design of Experiments** and **Response Surface Modeling**. Key goals include **Efficiency** in minimizing simulations, **Flexibility** in integrating simulators and optimization techniques, and implementation using **open-source and proprietary tools**.
- The work proposes a **Multi-Objective Design Space Exploration (DSE) framework** to customize MP-SoC architectures evaluating a set of metrics.
 - The DSE framework is simulation-based and focusing on:
 - **Design of Experiments**
 - **Response Surface Modeling**
 - **Efficiency** of design space exploration in terms of minimizing the number of simulations
 - **Flexible** DSE methodology: easy plug-in of system-level simulators and optimization techniques
 - Framework implemented in a set of **open-source and proprietary tools**
- 20 Cristina SILVANO - Politecnico di Milano (ITALY)

MULTI³

Design Flow Integration based on XML interface between design tools

- Definition of the specification for the design flow integration: The formal specification of the tool interface, based on the XML standard, is of fundamental importance for granting the seamless integration of design tools into a common design environment.

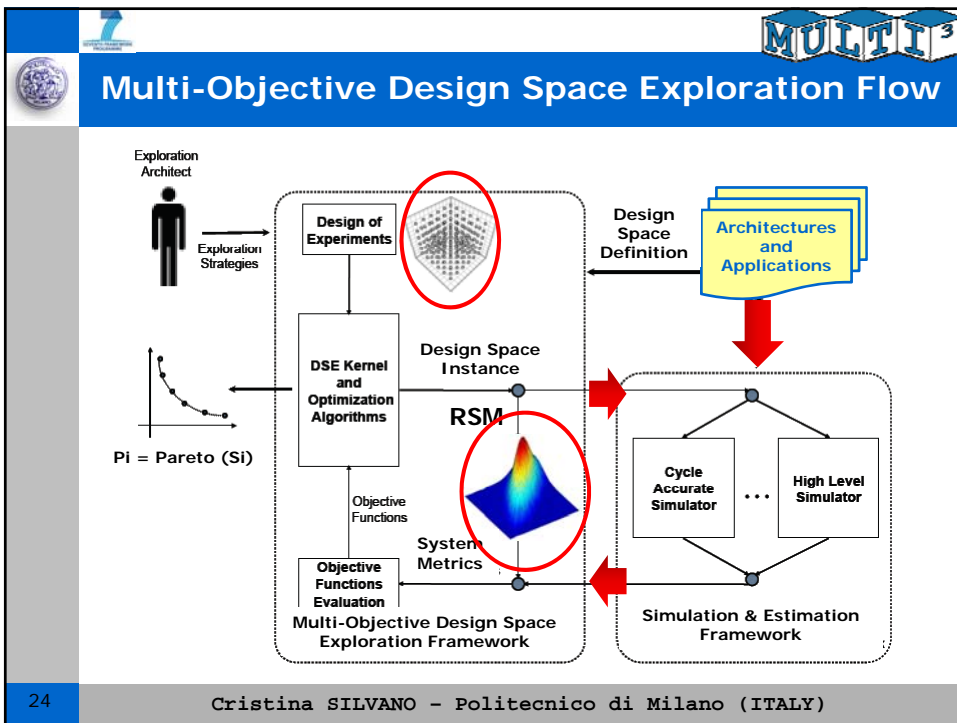
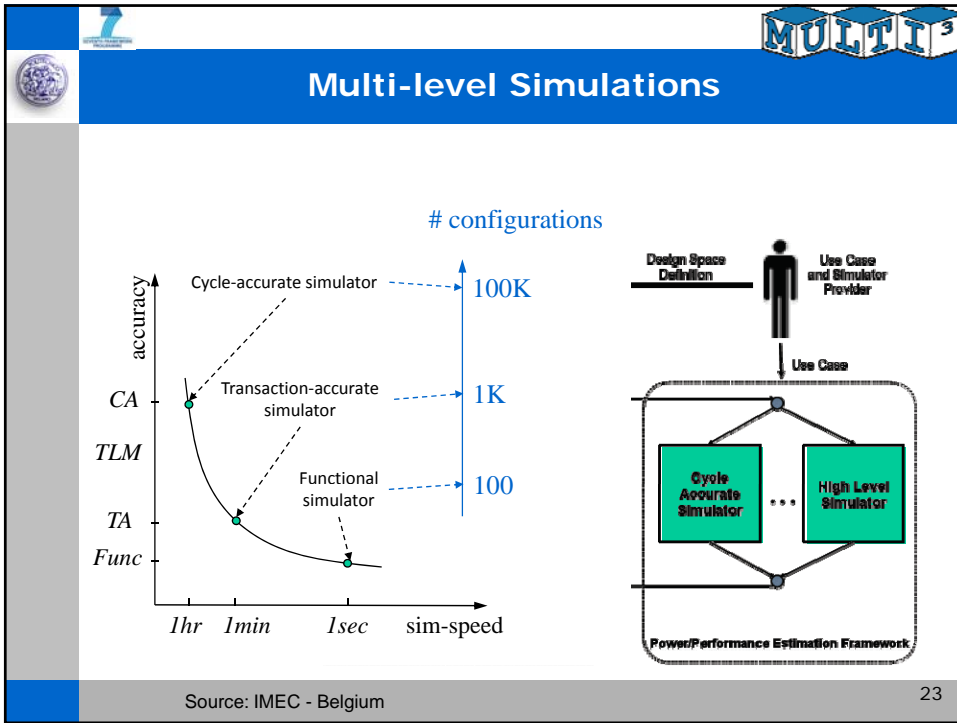
5 / 28 / 2009

MULTI³

Multi-Objective Design Space Exploration Flow

22

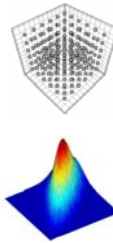
Cristina SILVANO - Politecnico di Milano (ITALY)



MULTI³

Multi-Objective Design Space Exploration

- MO-DSE framework based on:
 - **Design of Experiments (DoEs):**
To identify the experimentation plan where the set of tunable design parameters can vary
 - **Response Surface Modeling (RSM):**
To use the set of data generated by DoE to obtain a response surface of the system behavior
- RSM based on two main phases:
 - During the **training phase**, known data (or training set) are used for tuning the RSM.
 - During the **prediction phase**, the RSM is used to predict the unknown system response.



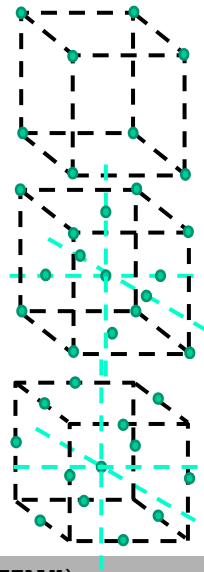
25

Cristina SILVANO - Politecnico di Milano (ITALY)

MULTI³

Design of Experiments

- Identifies the planning of experimentation campaign where the set of tunable design parameters can vary
- Specifies the **layout**: how to select the design points in the design space
- **Four DoE techniques have been applied:**
 - **Random**
 - **Full Factorial**
 - **Central Composite**
 - **Box Behnken**



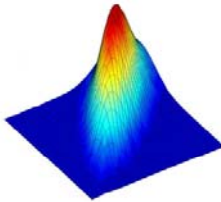
26

Cristina SILVANO - Politecnico di Milano (ITALY)

MULTI³

Response Surface Modeling

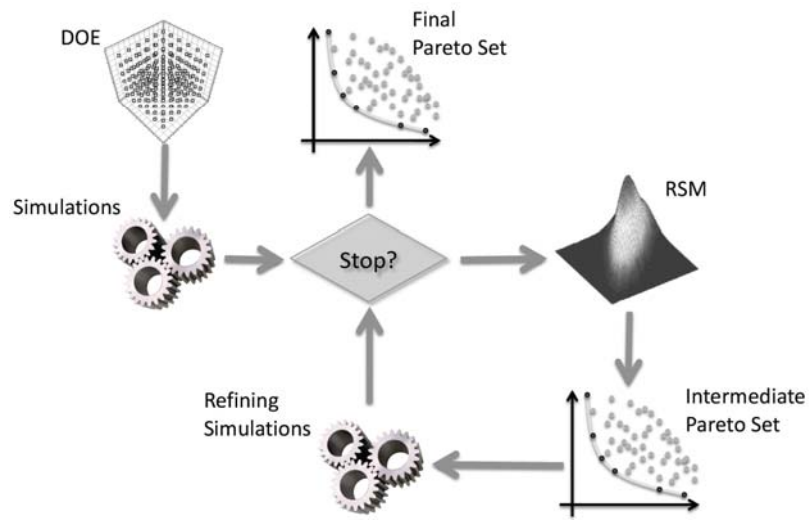
- RSM techniques are used to define an analytical dependence between design parameters and one or more response variables.
- To use the set of data generated by DoE to obtain a response model of the system behavior to forecast unknown system response.
- Four DoE techniques have been applied:
 - RSM based on Linear Regression
 - RSM based on Shepard's Interpolation
 - RSM based on Artificial Neural Networks
 - Three-layer fully-connected feed-forward ANNs
 - RSM based on Radial Basis Functions



27 Cristina SILVANO - Politecnico di Milano (ITALY)

MULTI³

RSM-Support Iterative Pareto Refinement



DOE

Simulations

Stop?



Refining Simulations

RSM

Intermediate Pareto Set

Final Pareto Set

28 Cristina SILVANO - Politecnico di Milano (ITALY)



Target MP-SoC Architecture

- MIPS-based shared memory MP-SoC with private caches
- Modeled with SESC simulator with power estimation support

Parameter	Min.	Max.
# Processors	2	16
Processor issue width.	1	8
L1 instruction cache size	2K	16K
L1 data cache size	2K	16K
L2 private cache size	32K	256K
L1 instruction cache assoc.	1w	8w
L1 data cache assoc.	1w	8w
L2 private cache assoc.	1w	8w
I/D/L2 block size	16	32

- Design space composed of 2^{17} design points (131 072)
- Four parallel applications {FFT, OCEAN, LU, RADIX} derived from SPLASH-2 benchmark suite for different data-sets

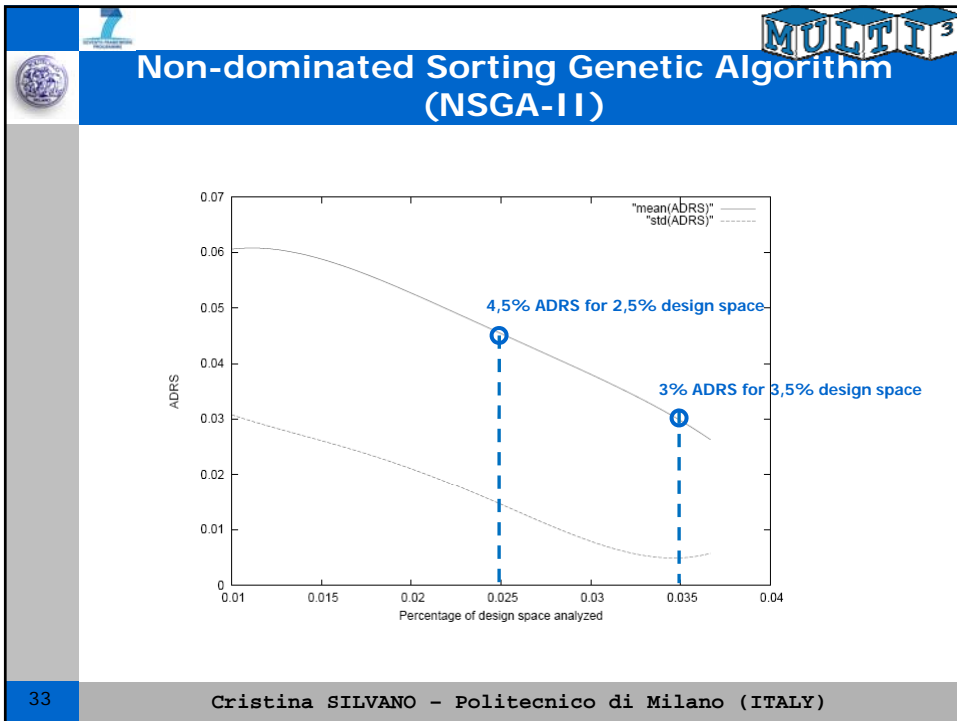
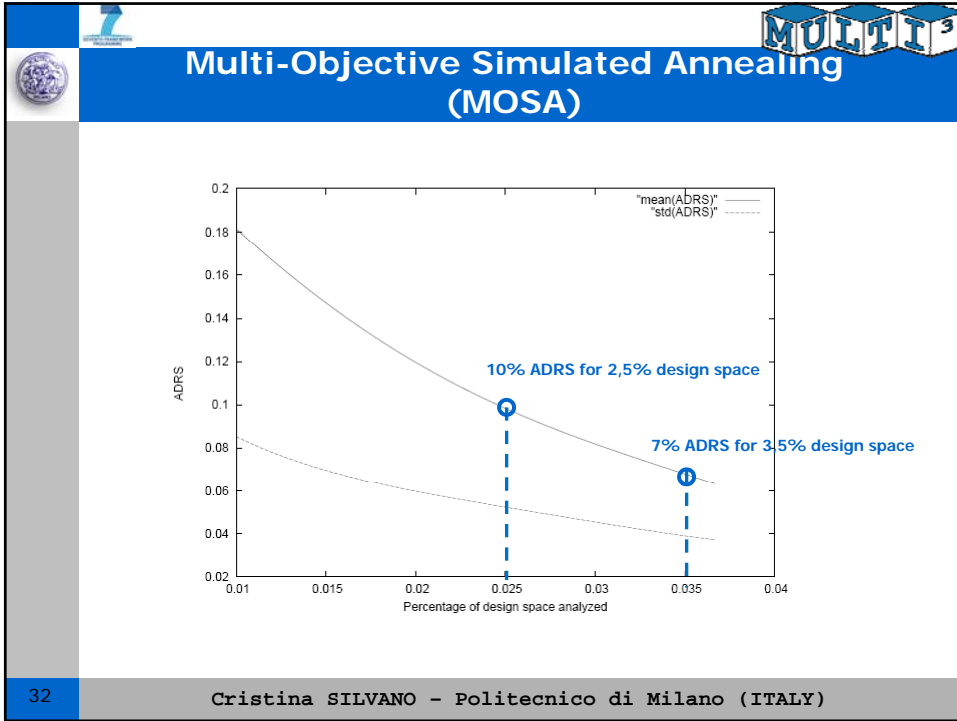
30 Cristina SILVANO - Politecnico di Milano (ITALY)

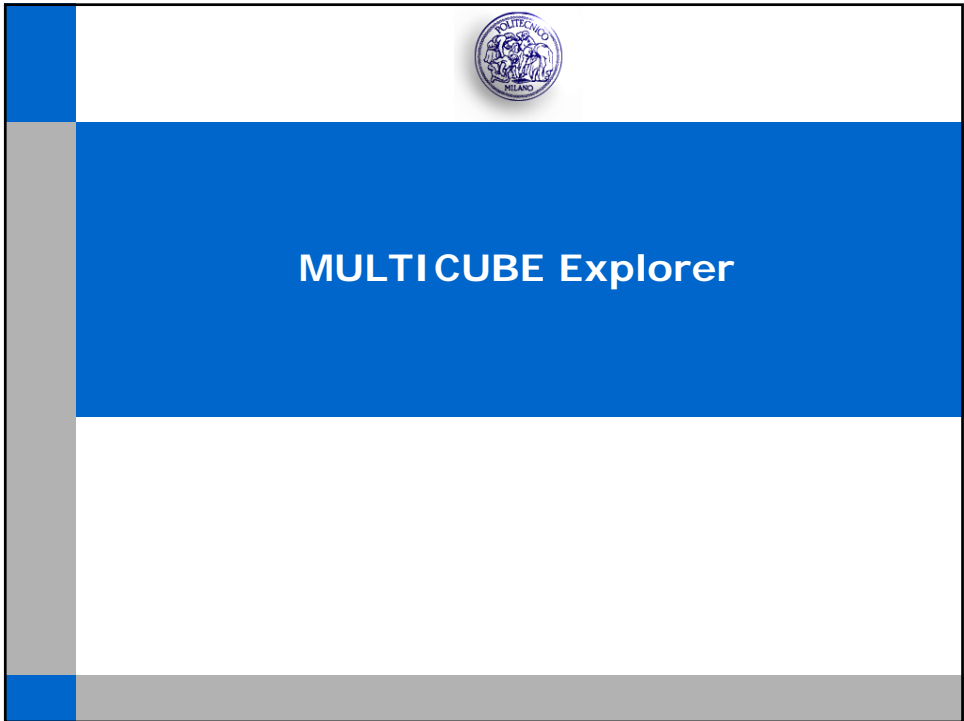
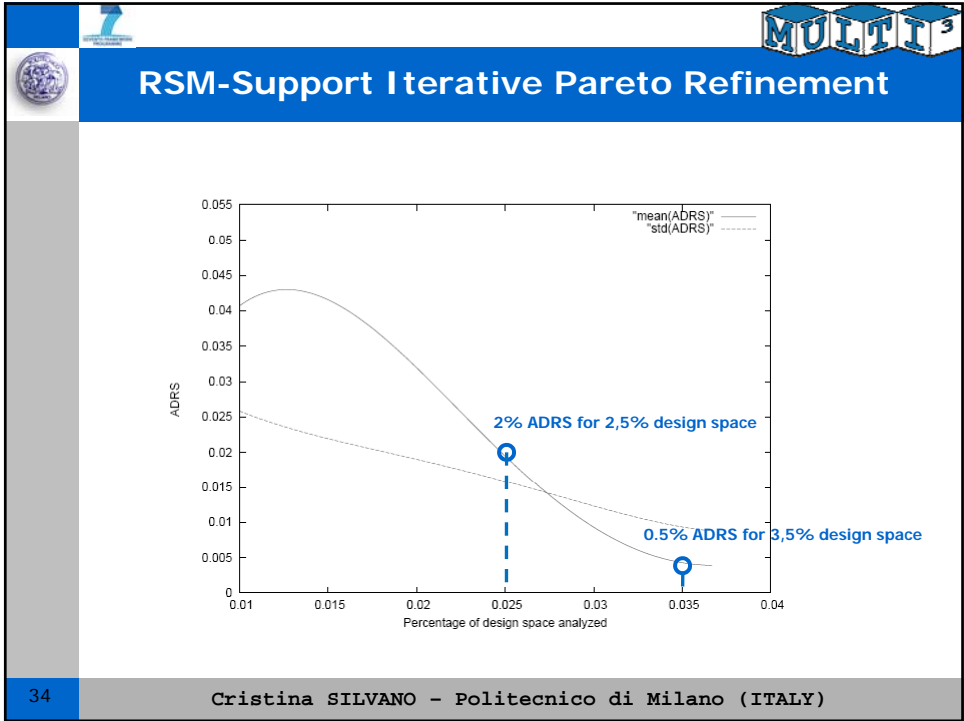



Experimental Results

- Target multi-objective optimization problem:
Minimization of *average execution time* and *average [mW per MIPS]* over the set of several application scenarios and subject to total cache size constraint
- Accuracy in terms of **Average Distance from Reference Set** to measure the distance between reference Pareto front and approximated Pareto front
 - **Lower ADRS, best approximated Pareto front**
- ADRS by varying the percentage of the design space analyzed from 1% to 3,5%
- Comparison with state-of-the-art heuristics:
 - **Multi-Objective Simulated Annealing (MOSA)**
 - **Non-dominated Sorting Genetic Algorithm (NSGA-II)**

31 Cristina SILVANO - Politecnico di Milano (ITALY)





MULTI³

MULTICUBE Explorer

- **Open-source prototype exploration framework (MULTICUBE Explorer):**
The tool enables a fast automatic optimization of parameterized system architectures towards a set of multiple objectives

www.multicube.eu

```

    graph LR
      UC[Use Case Simulator] -- "XML System Metrics" --> ME[Multicube Explorer]
      ME -- "XML System Configuration" --> UC
      M3S[M3 Explorer Shell] --> M3K[M3 Explorer Kernel]
      XMLDF[XML Design Space Definition File] --> M3K
      M3K --> DE[Design of Experiments]
      M3K --> OA[Optimization Algorithm]
      M3K --> UC2[Use Case Simulator]
      DE --> OA
      OA --> AD[(Architecture Database)]
      UC2 --> AD
  
```

5 / 28 / 2009

MULTI³

Command line interface

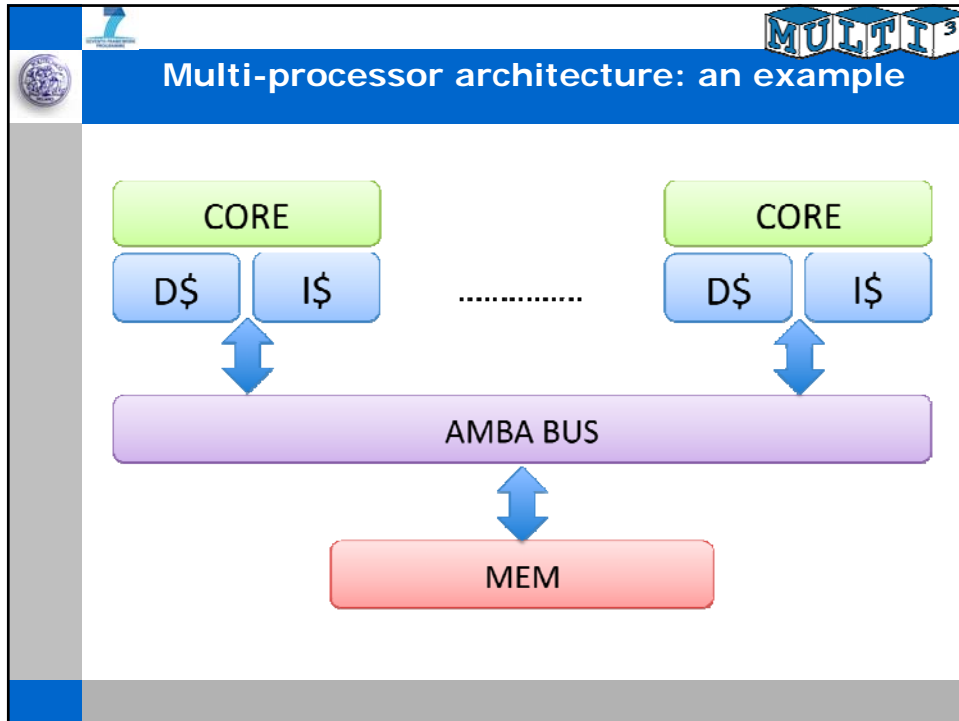
```

    Default
    hbomb:~/projects/trunk/m3explorer/build> ./image/bin/m3explorer

    M3EXPLORER

    Multicube Explorer - Version snapshot_200309
    Send bug reports to zaccaria@elet.polimi.it, gpalermo@elet.polimi.it
    --

    m3_shell>
  
```



Example of XML design space and metrics

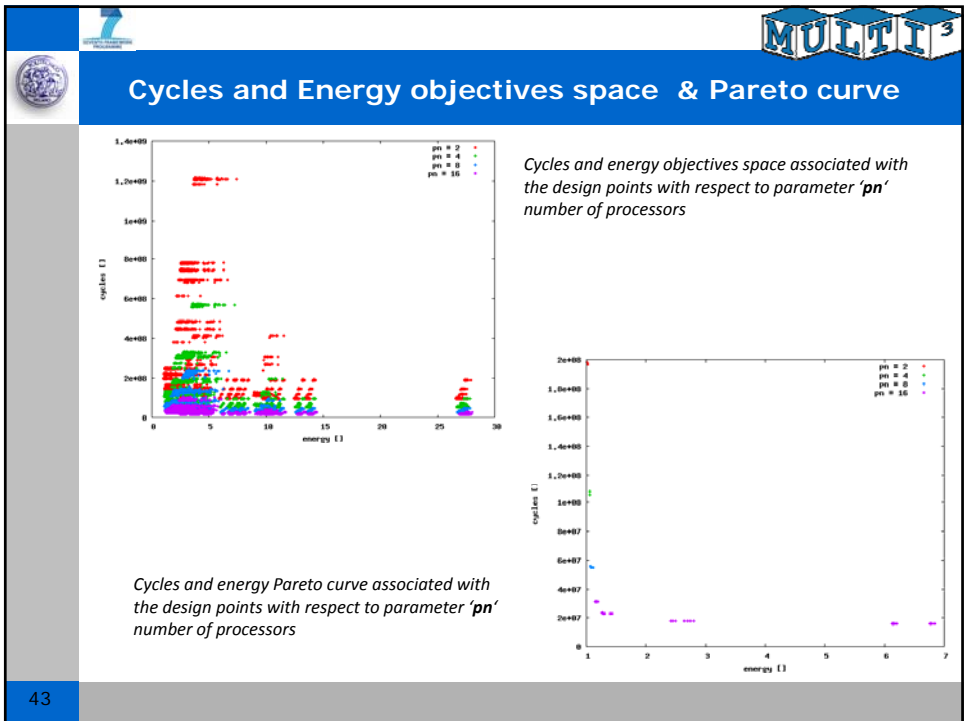
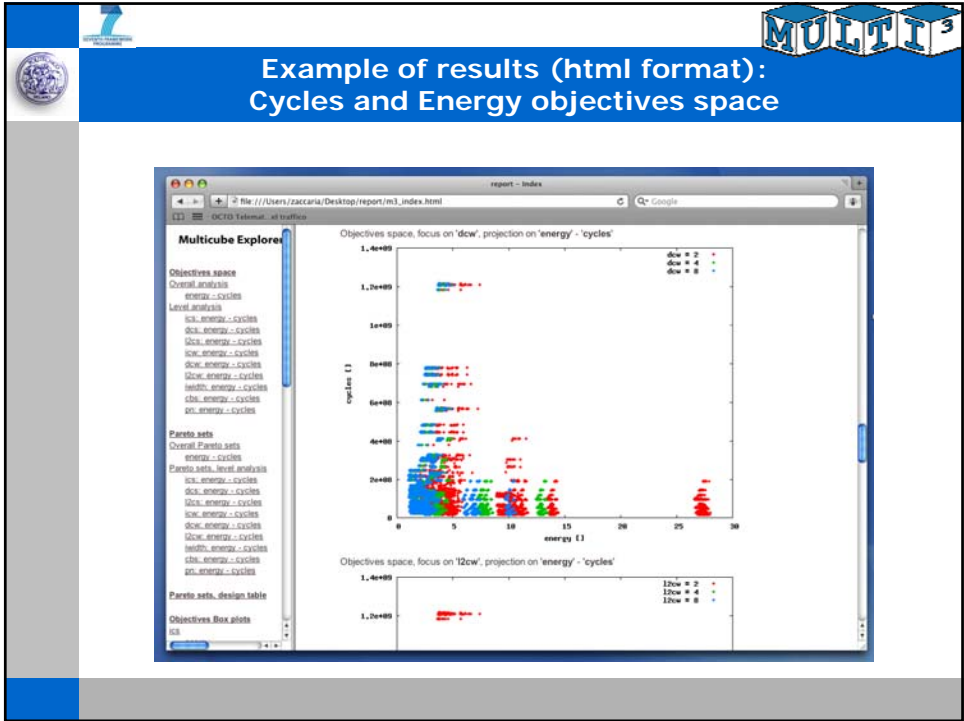
```

<?xml version="1.0" encoding="UTF-8"?>
<design_space xmlns="http://www.multicube.eu/" version="1.3">
  <simulator>
    <simulator_executable
      path="/home/demo/tools/multicube-scope/qcif_example_v2/scope_example/script.sh" />
    </simulator>
    <parameters>
      <parameter name="num_cpus" type="integer" min="2" max="8" />
      <parameter name="icache_size" type="exp2" min="4096" max="32768"/>
      <parameter name="freq" type="integer" min="40" max="200" step="40"/>
    </parameters>
    <rules>
      <rule>
        <not-equal>
          <parameter name="num_cpus"/> <constant value="3"/>
        </not-equal>
      </rule>
    </rules>
    <system_metrics>
      <system_metric name="latency" type="float" unit="Second" />
      <system_metric name="instruction_count" type="float" unit="Instruction"/>
      <system_metric name="power_consumption" type="float" unit="W" />
    </system_metrics>
  </design_space>

```

The XML code is annotated with callouts for its different sections:

- Path:** Points to the `<simulator_executable path="..." />` element.
- Parameters:** Points to the `<parameters>` block containing three `<parameter>` elements.
- Rules:** Points to the `<rules>` block containing a `<rule>` with a `<not-equal>` constraint.
- Metrics:** Points to the `<system_metrics>` block containing three `<system_metric>` elements.





Conclusions

- An automatic design space exploration methodology has been proposed leveraging Design of Experiments and Response Surface Modeling techniques
- The proposed methodology makes automatic exploration of CMP architectures more feasible
- The proposed approach can be easily combined with fast simulation techniques
- Future work: Joint architecture and compiler spaces to be explored
- This work is part of the ICT-FP7 EU project MULTICUBE

www.multicube.eu