POLITECNICO DI MILANO

Design Space and Application Autotuning for Runtime Adaptivity in Multicore Architectures

POLITECNICO MILANO 1863

Cristina Silvano
Politecnico di Milano
cristina.silvano@polimi.it

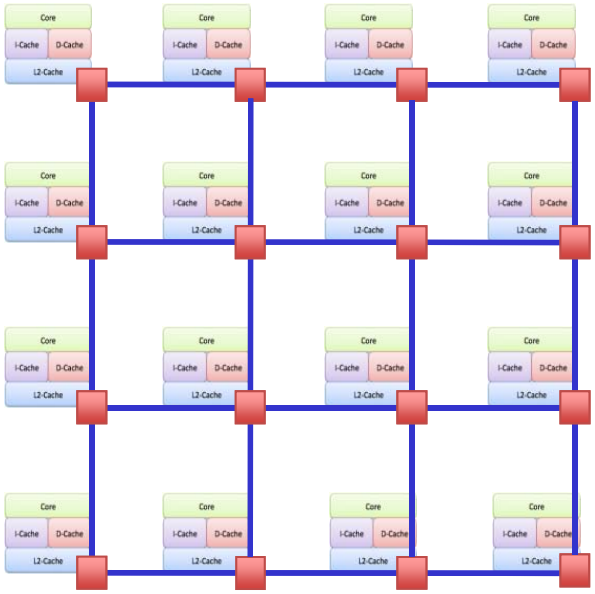
Energy efficiency underlies all markets

- **Energy efficiency** is of paramount importance for all application markets (automotive, consumer, mobile, healthcare and beyond) and target systems spanning from sensors, cyber-physical systems, embedded systems up to servers and HPC systems.

The design space in the age of multi/many-cores

- A wide range of architecture parameters must be tuned to find the best system configuration.
- Design space of the target architecture A should consider all possible configurations of each parameters p_i :

$$A = S_{p1} \; x \; S_{p2} \; x \; .... \; x \; S_{pn}$$

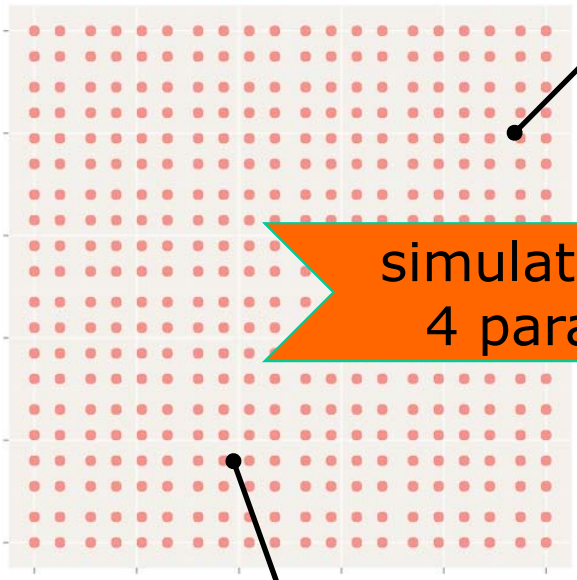


Parameter	Min	Max
#Processors	2	16
#Threads	1	4
L1 I$ size	2K	16K
L1 D$ size	2K	16K
L1 I$ associativity	1w	8w
L1 D$ associativity	1w	8w
L2 $ size	32K	256K
L2 $ associativity	1w	8w

Large design space: 2^{18} (262 144) design points

Full search simulation time

FULL SEARCH
Can become quickly unfeasible



~10 minutes
per simulation*

simulation time on
4 parallel cores

10 Years

262 144 design points
x
8 data sets
=
2 097 152 simulations

* Using a cycle-accurate simulator

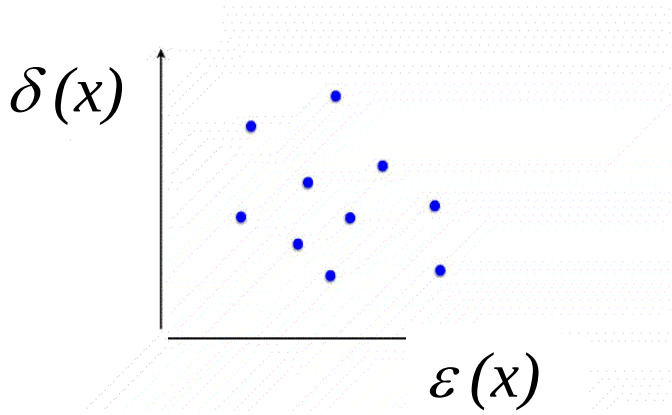
The multi-objective optimisation problem

- **Objective function:** To minimize both energy $\varepsilon(x)$ and execution time $\delta(x)$ of the target application on system configurations *x:*

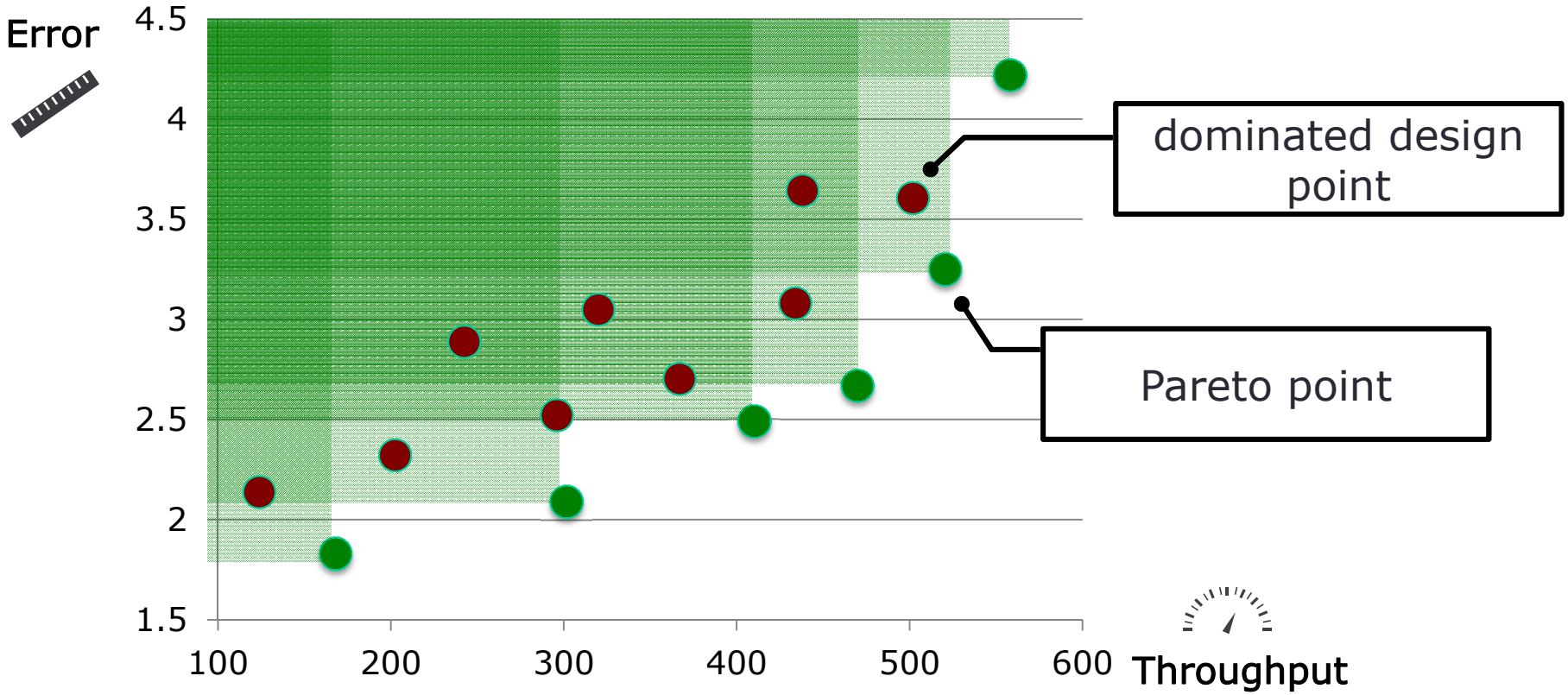
$$\min_{x \in X} \omega(x), \ \omega(x) = \begin{bmatrix} \epsilon(x) \\ \delta(x) \end{bmatrix}$$

where X is the design space.

- The solution is a set of tradeoff configurations $X_p \subseteq X$ known as Pareto set



Multi-objective exploration: Pareto Points

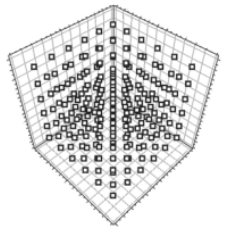


Multi-Objective Exploration: **best designs are not unique**.
Pareto points provide tradeoffs with respect to the multiple objectives

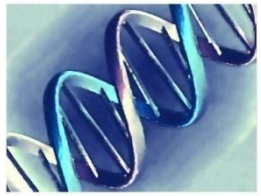
MULTICUBE Explorer

www.multicube.eu

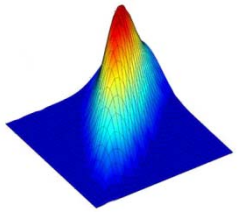
1. **Design of Experiments (DoEs):**
   To identify the experimentation plan: how to select the design points in the design space to be simulated: random, full factorial, central composite design.



2. **Optimisation Algorithms:**
   Meta-heuristics methods inspired by analogies with physics, or with biology to solve multi-objective optimization problems: simulated annealing, genetic algorithms, evolutionary strategies, etc.

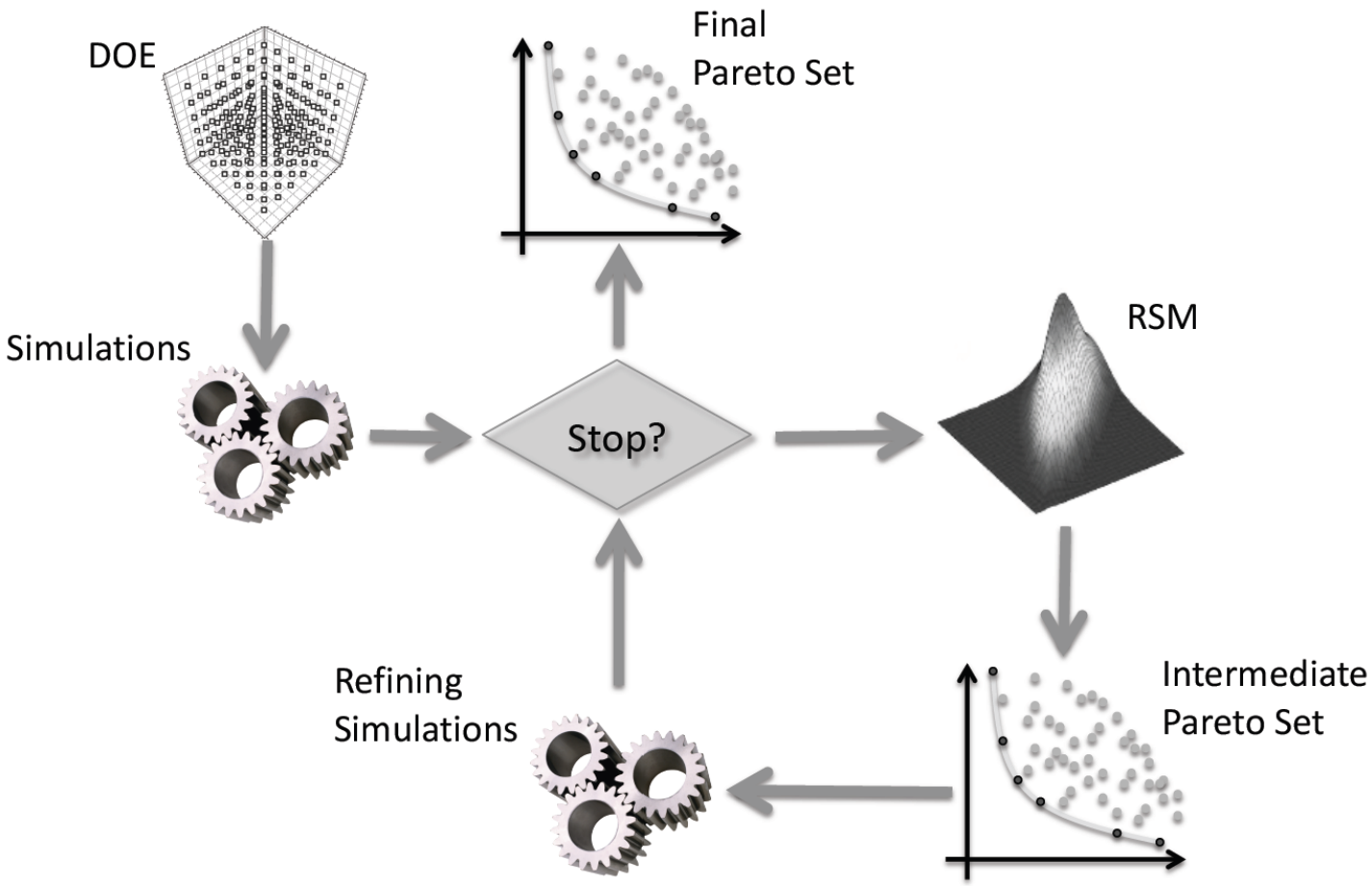


3. **Response Surface Modeling (RSM):**
   To use the set of simulated points to obtain an analytical model of the system behavior: linear regression, spline interpolation, artificial neural network, etc.



How to combine DoEs and Response Surface Models?

➢ ReSPIR: Response Surface-based Pareto Iterative Refinement

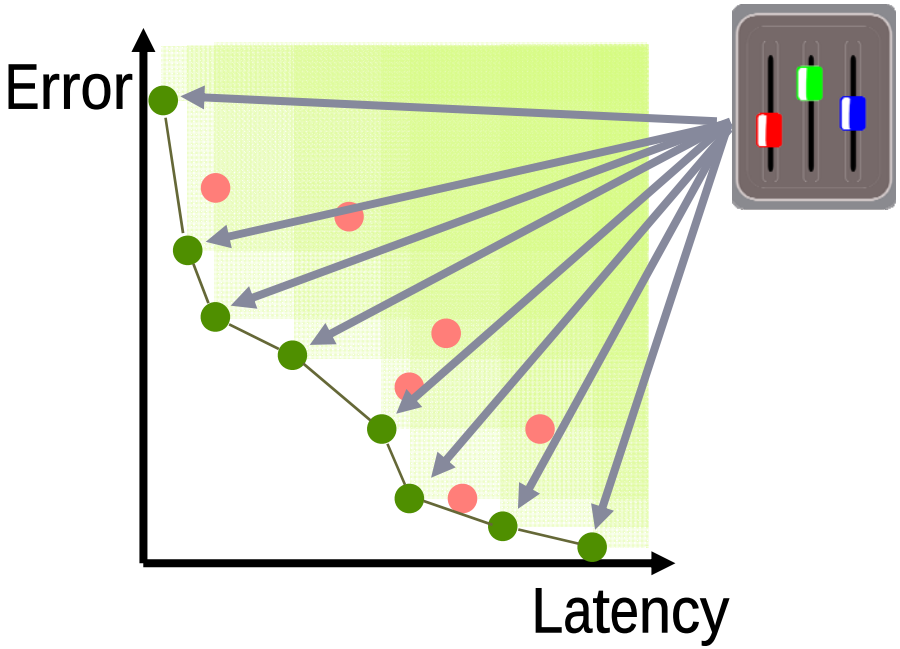


"ReSPIR: A Response Surface-Based Pareto Iterative Refinement for Application-Specific Design Space Exploration", Palermo, G.; Silvano, C.; Zaccaria, V., *IEEE Trans. on CAD of Integrated Circuits and Systems,* Vol. 28, Issue 12, Dec. 2009 Page(s): 1816 – 1829

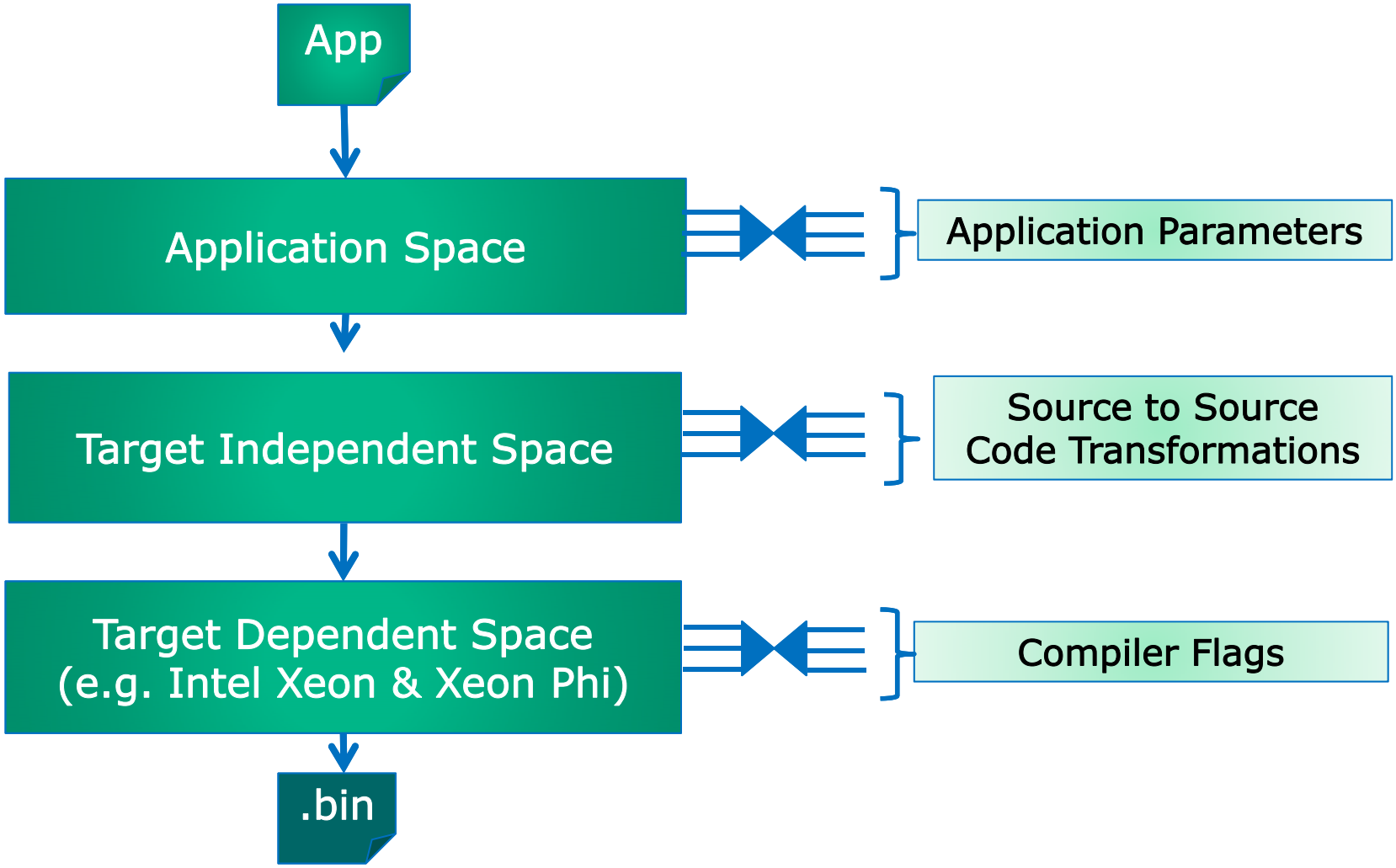
APPLICATION AUTOTUNING

mARGOt Application Autotuning

- **Key idea** is that most of the applications are configurable at **design-time** in terms of a set of tunable parameters, code transformations and code variants **(application knobs)** to trade-off accuracy and latency



Tunable Applications: Software Knobs

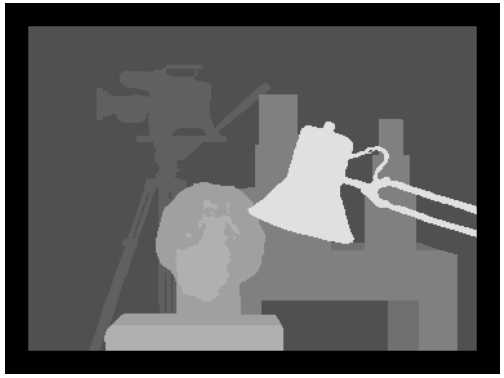


Quality of Results: Pixel Disparity Error

Left camera

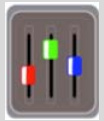
Right camera

stereo-matching



Reference disparity map

Application Knobs



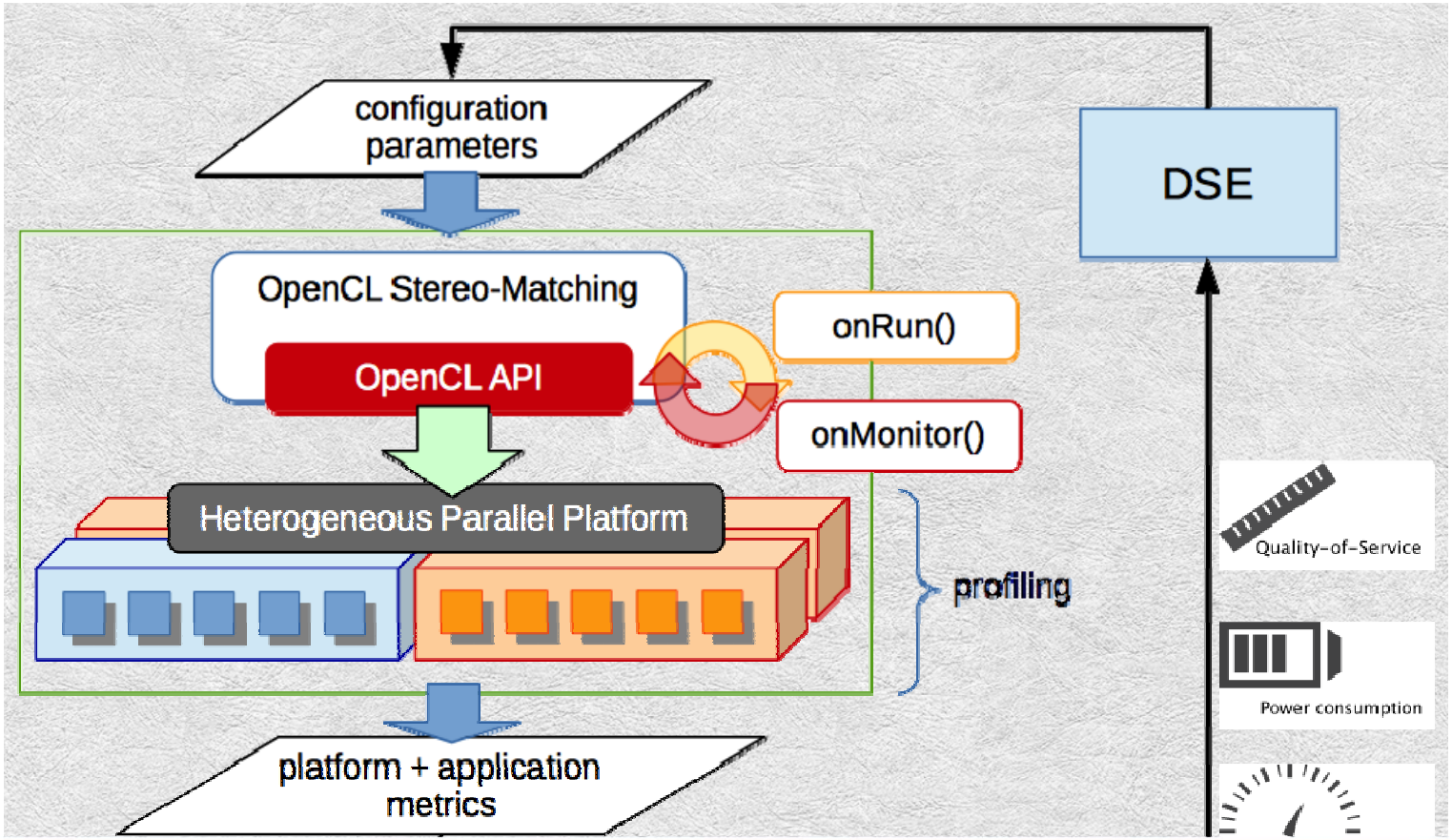
Disparity Error

1

2

3

Design-time optimization of OpenCL applications



E. Paone, G. Palermo, V. Zaccaria, C. Silvano, D. Melpignano, G. Haugou and T. Lepley, "An Exploration Methodology for a Customizable OpenCL Stereo-Matching Application Targeted to an Industrial Multi-Cluster Architecture", in Proc. CODES+ISSS 2012.

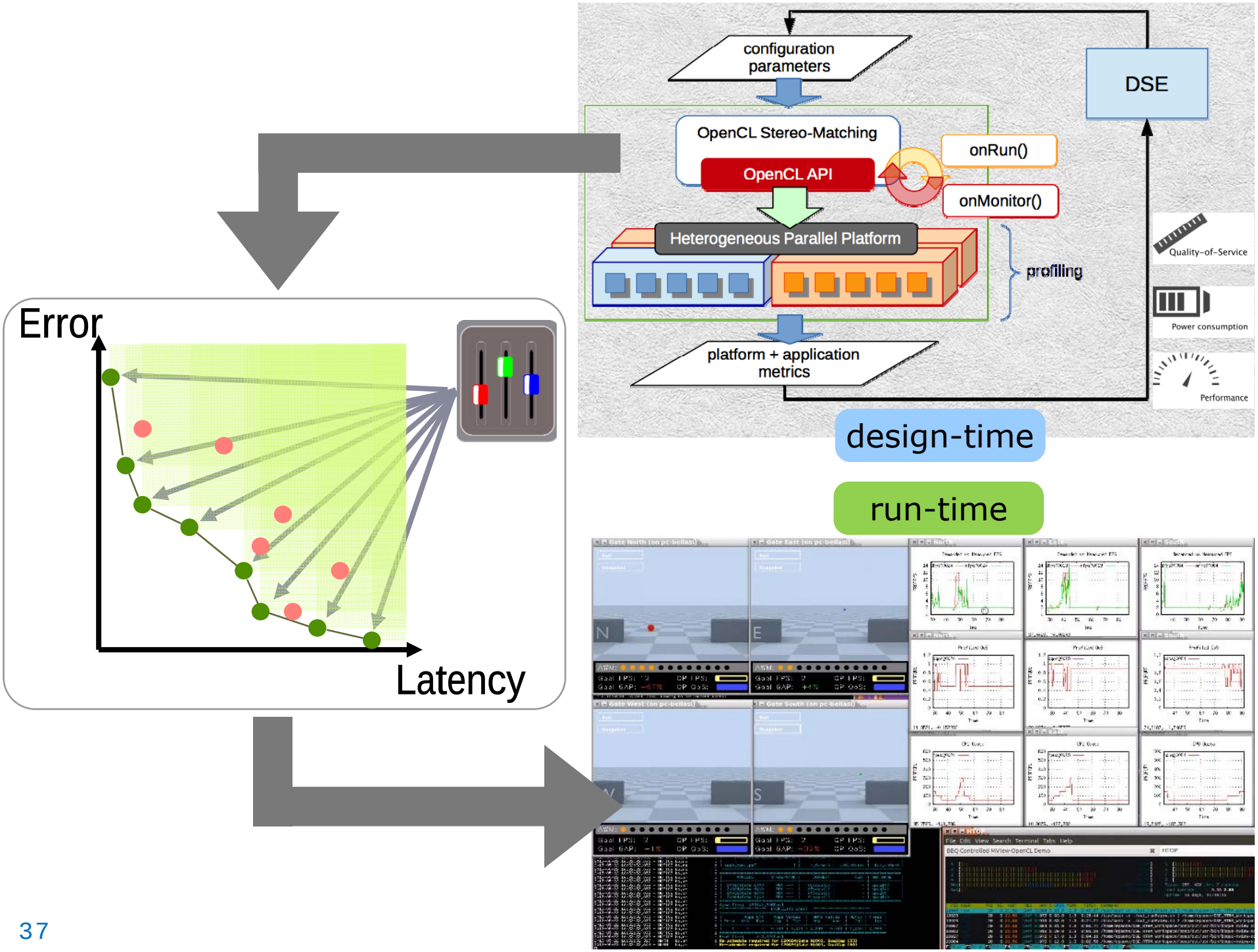
E. Paone, F. Robino, G. Palermo, V. Zaccaria, I. Sander and C. Silvano, *"Customization of OpenCL Applications for Efficient Task Mapping under Heterogeneous Platform Constraints"*, In Proceedings of DATE 2015.

mARGOt Application Autotuning

- **mARGOt** is a **light-weight application autotuning framework** for many-core platforms in an adaptive multi-application environment.
  - Combination of **design-time** and **run-time** techniques to create an effective way of *"self-aware"* computing with **limited runtime overhead**
  - **Adaptivity** to **runtime** adjust the application "behavior" (configuration) to the changing operating conditions, usage contexts and resource availability
  - **Approximate computing** to trading off accuracy and throughput while the output just needs to be "good enough"
  - **Orthogonality** between application autotuning and runtime

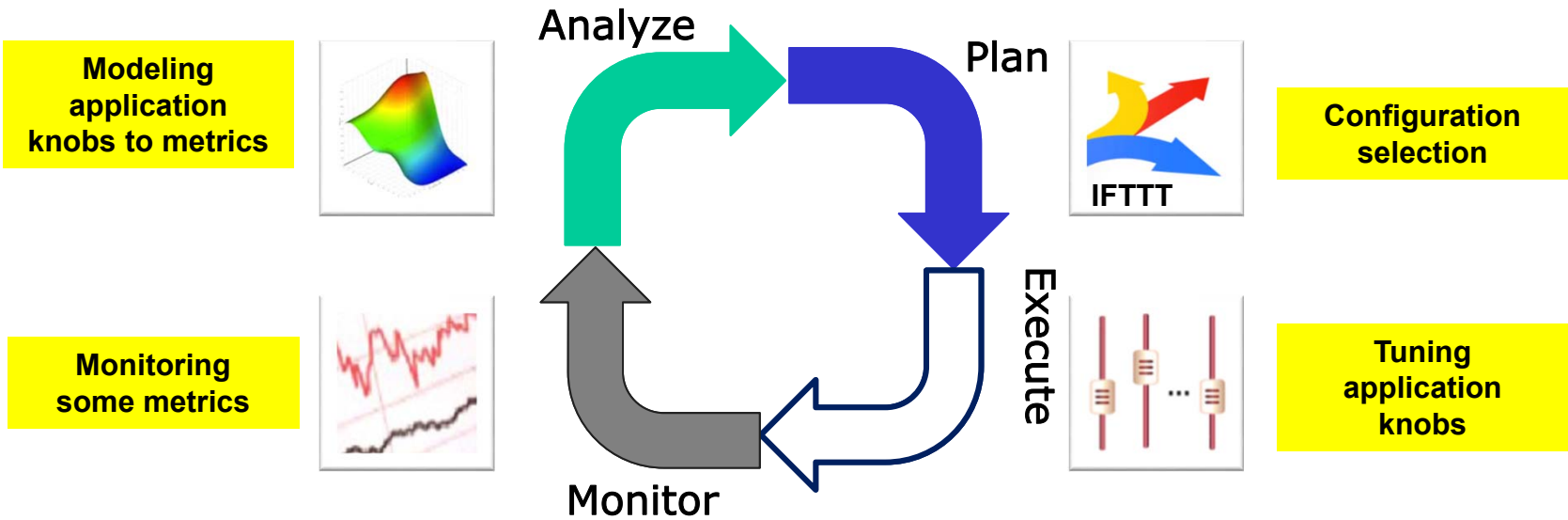
E. Paone, D. Gadioli, G. Palermo, V. Zaccaria, C. Silvano. "Evaluating Orthogonality between Application Auto-Tuning and Run-Time Resource Management for Adaptive OpenCL Applications", Proc. of IEEE ASAP 2014.

D. Gadioli, G. Palermo, C. Silvano, "Application Autotuning to Support Runtime Adaptivity in Multicore Architectures", in Proc. IEEE SAMOS 2015.

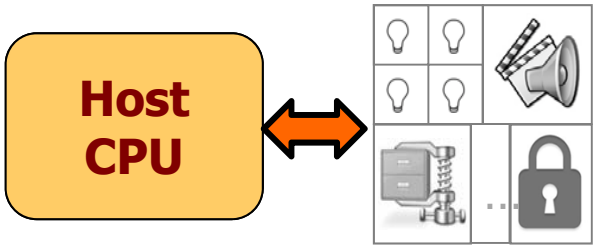

configuration parameters
DSE
OpenCL Stereo-Matching
OpenCL API
onRun()
onMonitor()
Heterogeneous Parallel Platform
profiling
Quality-of-Service
Power consumption
Performance
platform + application metrics
design-time
run-time
Error
Latency

Runtime feedback loop

- **Application autotuning** enables self-optimization capabilities based on **Monitor-Analyze-Plan-Execute (MAPE) feedback loop***



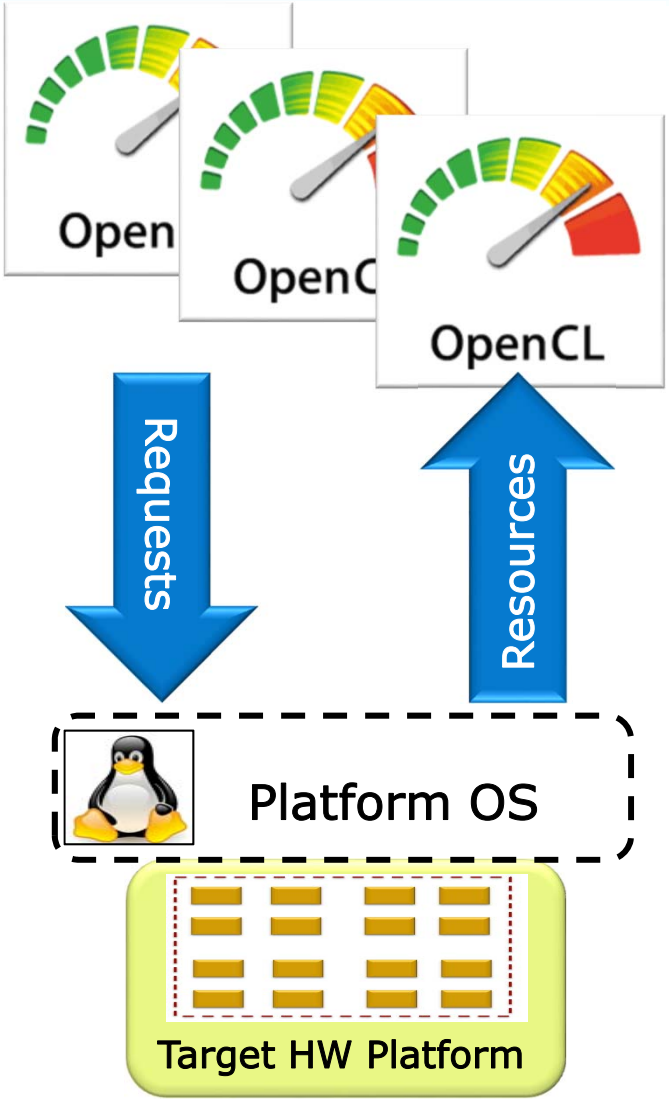
Heterogeneous Many-Core Platform



* J.O. Kephart, D.M. Chess, "The vision of autonomic computing," Computer, 2003

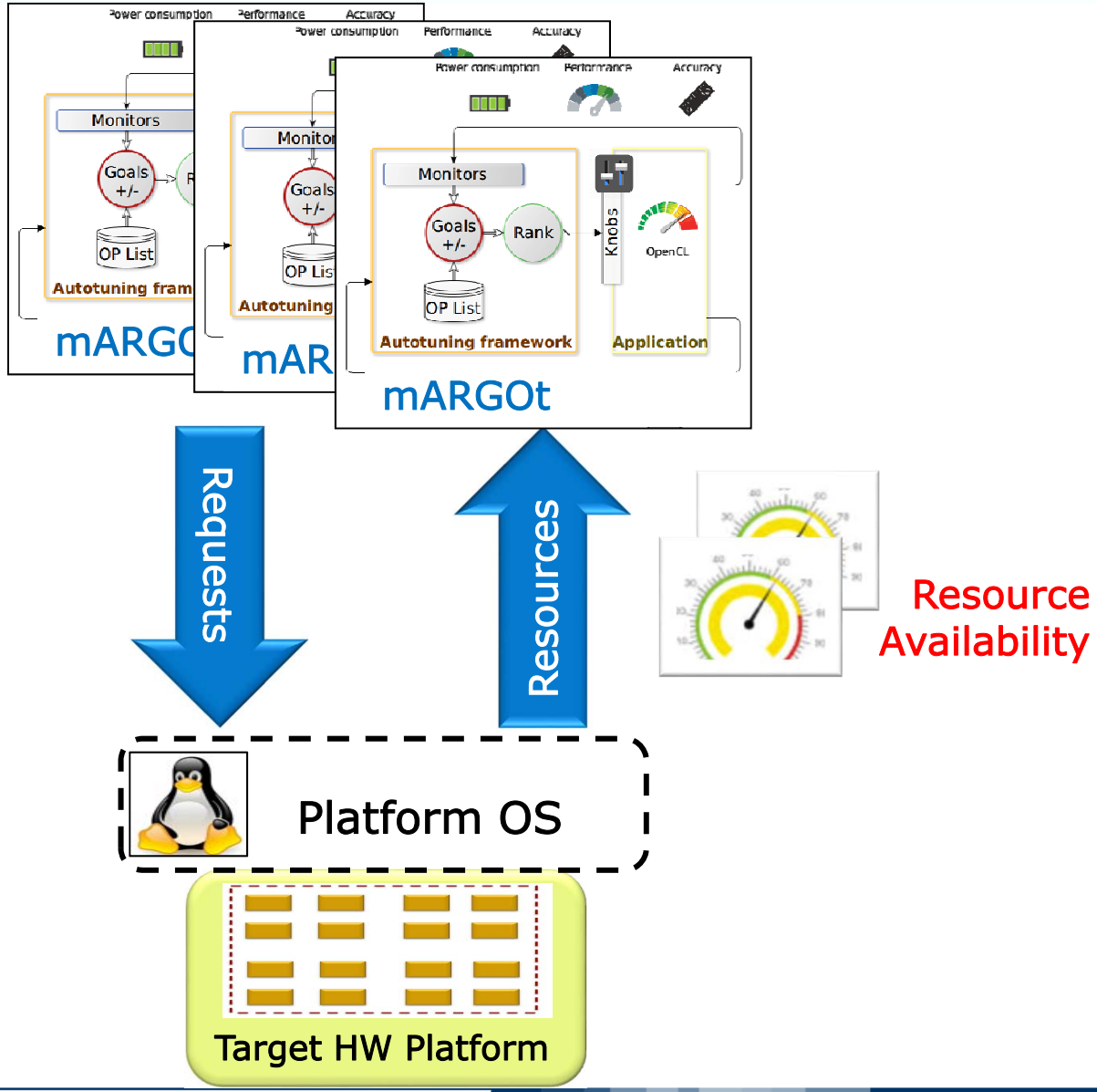
APPLICATION AUTOTUNING AND RUNTIME RESOURCE MANAGEMENT

Orthogonality Concept: App Autotuning & RTRM

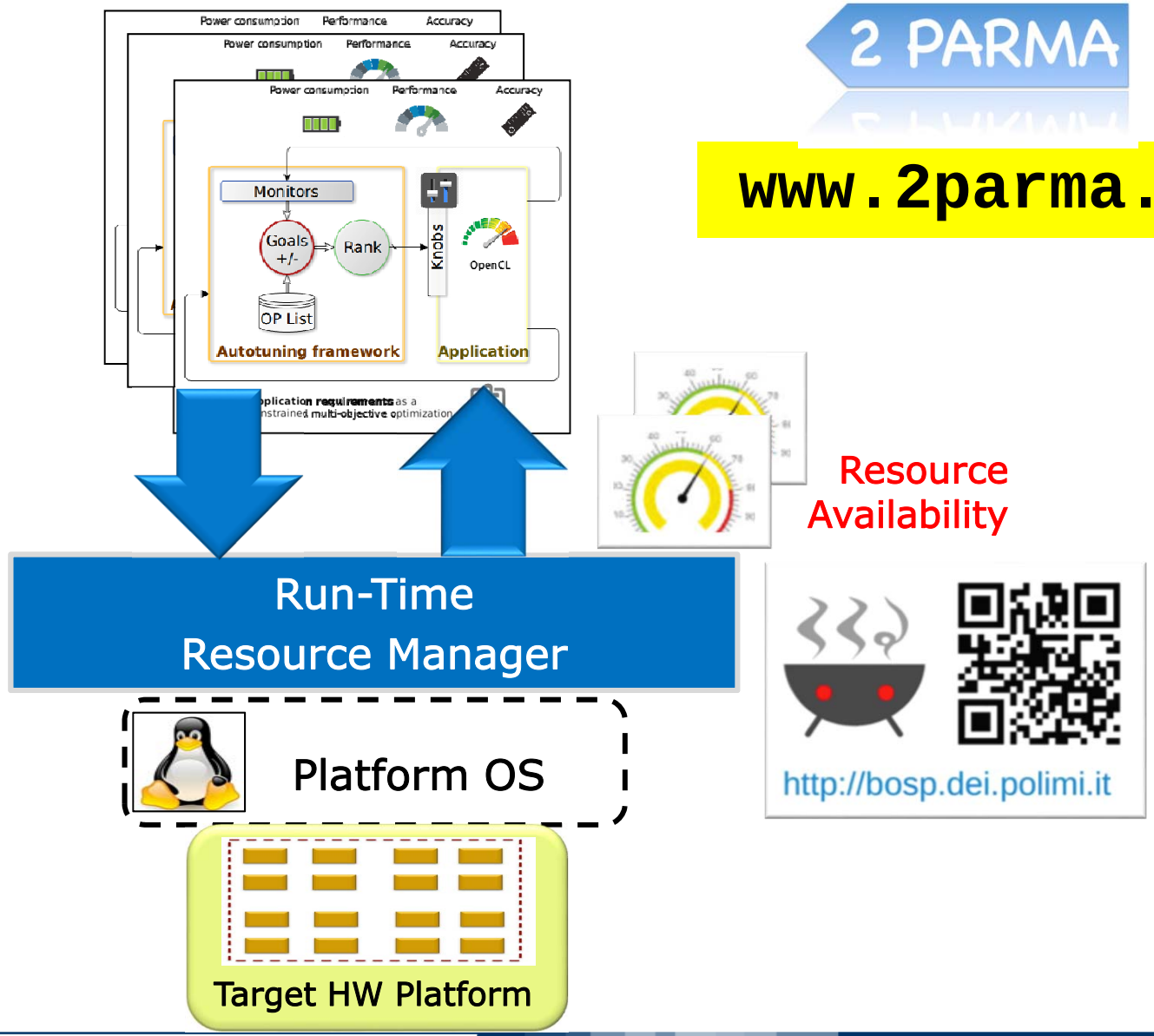


G. Massari, E. Paone, P. Bellasi, G. Palermo, V. Zaccaria, W. Fornaciari, and C. Silvano. "Combining Application Adaptivity and System-wide Resource Management on Multi-Core Platforms", Proc. of IC-SAMOS 2014.

Orthogonality Concept: App Autotuning & RTRM



Orthogonality Concept: App Autotuning & RTRM



ANTAREX

European Commission | Horizon 2020 European Union funding for Research & Innovation

http://www.antarex-project.eu/

Autotuning aNd adapTivity AppRoach for Energy efficient eXascale HPC systems

Prof. Cristina Silvano, Politecnico di Milano
cristina.silvano@polimi.it

ANTAREX Motivations

Context:

- To meet DARPA's target of **20MW** of **Exascale supercomputers**
  - Current supercomputers must achieve an energy efficiency "quantum leap"
  - **10x** energy efficiency from **5 to 50 GFlops/W**
- Next generation supercomputers need to be coupled with a radically **new software stack** capable of exploiting the benefits offered by heterogeneity to meet the scalability and energy efficiency required by the Exascale era.

ANTAREX Motivations

Context:

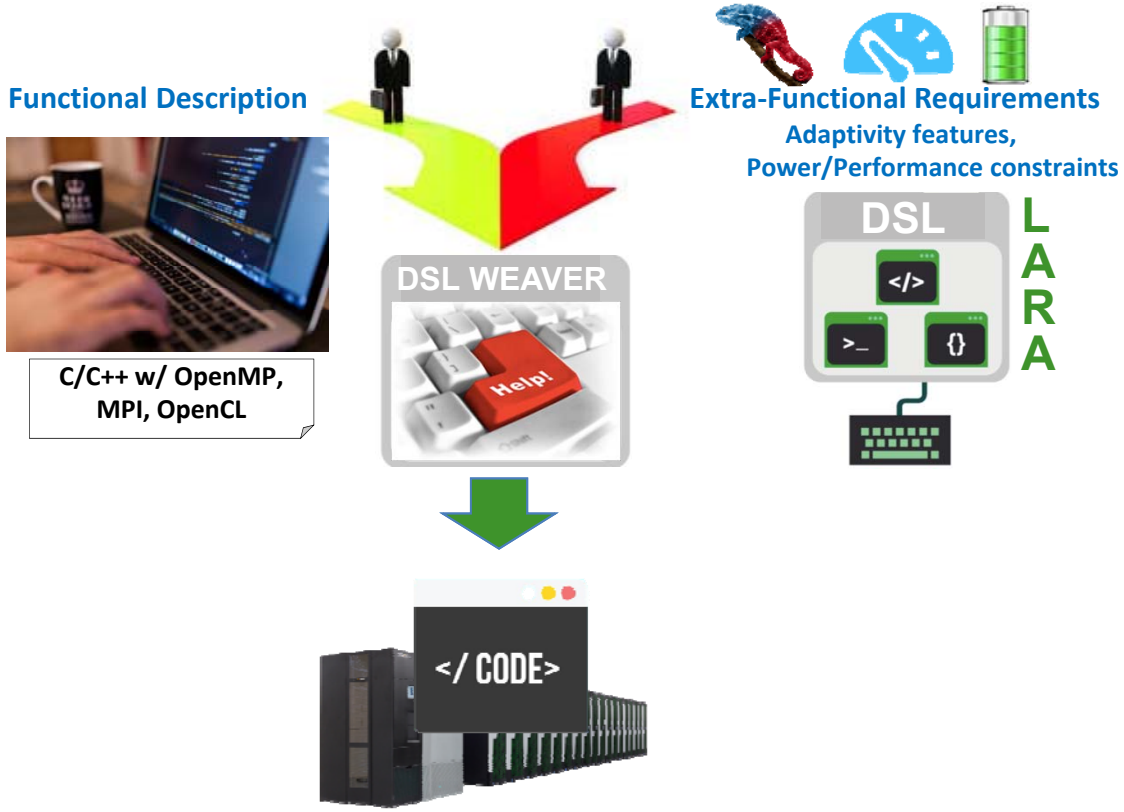
ANTAREX wants to provide a breakthrough approach

- to express by a Domain Specific Language the application self-adaptivity and non-functional requirements
- to runtime monitor, manage and autotune applications

for green HPC systems up to the Exascale level

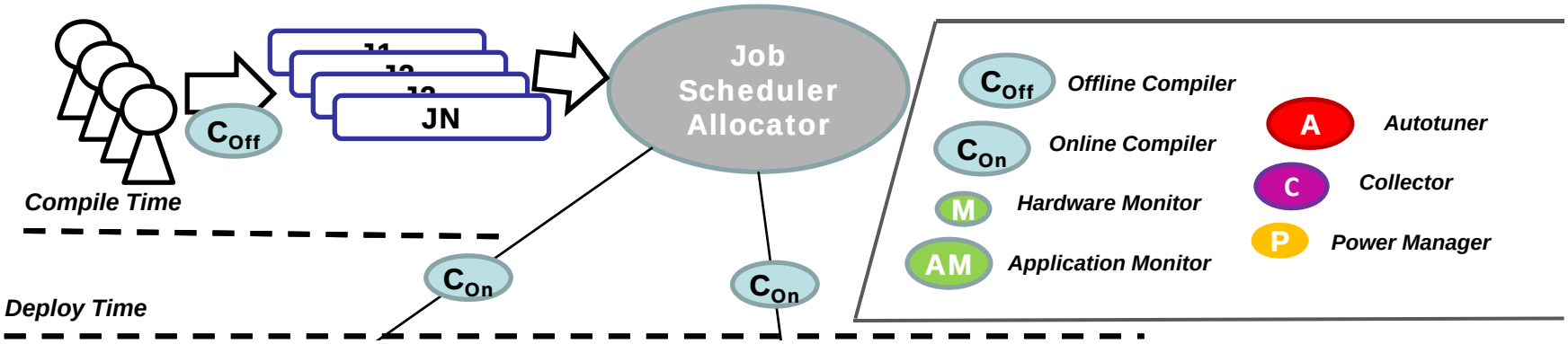
benefits offered by heterogeneity to meet the scalability and energy efficiency required by the Exascale era.

LARA DSL for Extra Functional Specifications

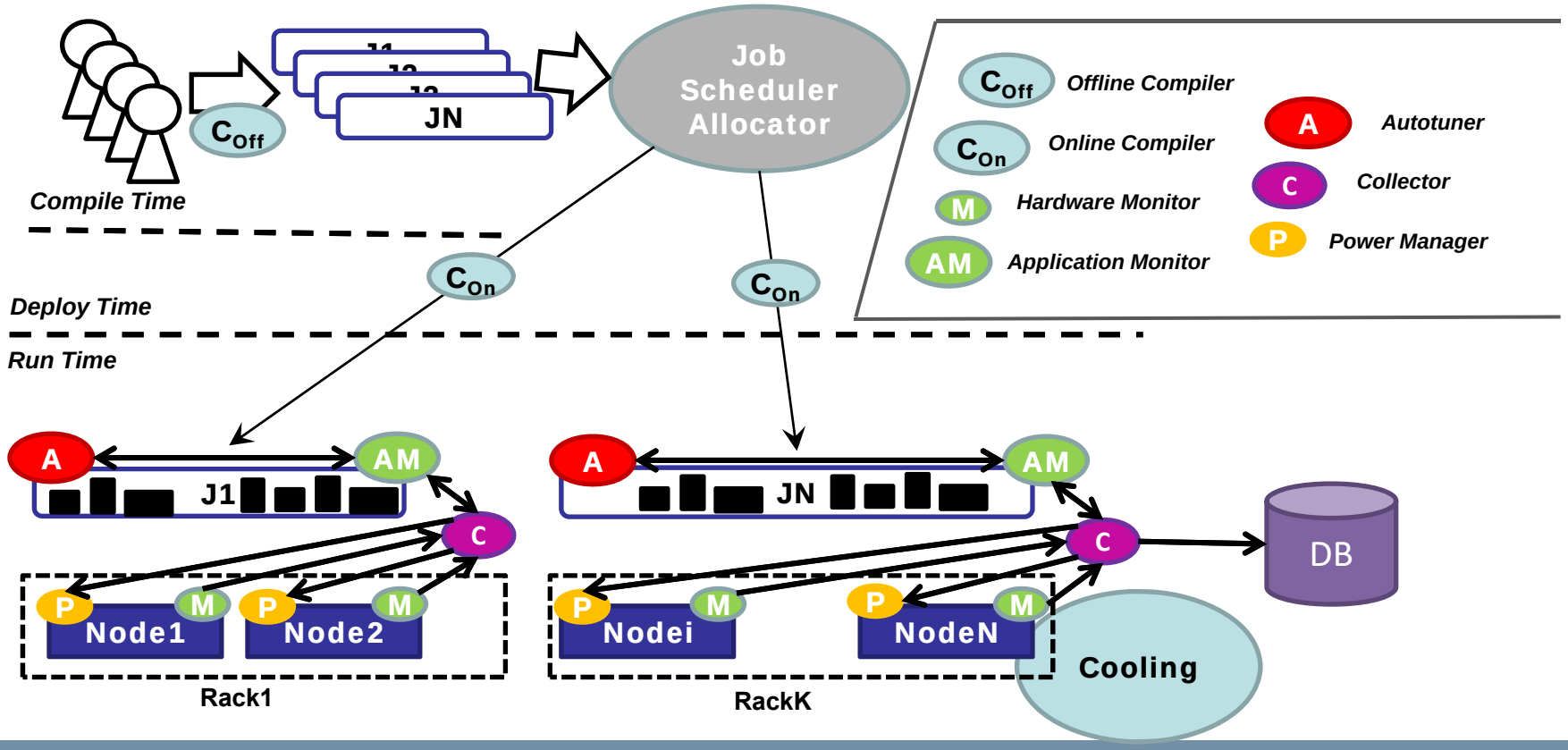


- To enable **separation of concerns:** functional and extra-functional descriptions are decoupled.
- Useful to express **strategies** for instrumentation and profiling
- Support for explore **compiler optimization space,** according to code and target architectures
- To enable more advanced **control** than using pragmas/directives/switches (e.g. application parameters and mapping control)

ANTAREX Runtime Framework

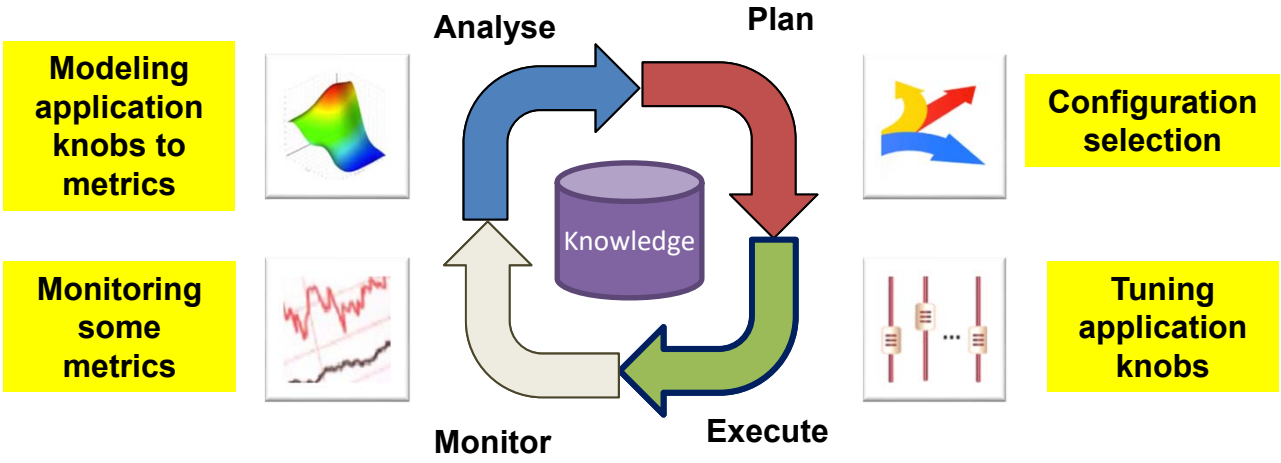


ANTAREX Runtime Framework



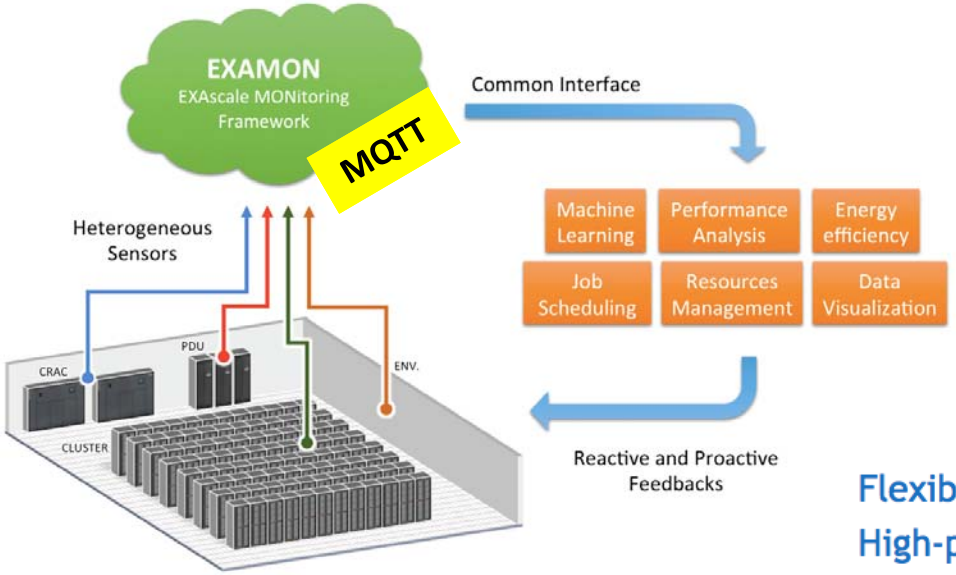
ANTAREX Application Autotuning Loop

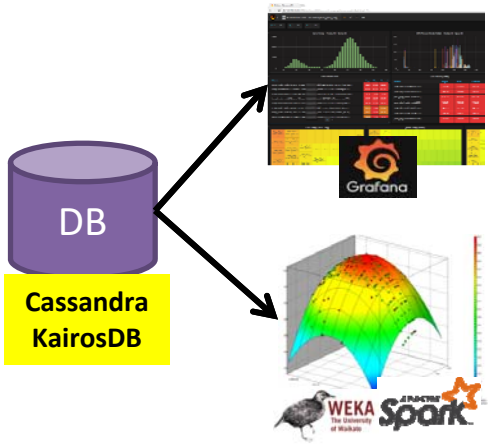
Autotuning framework implements a ***MAPE-K loop*** *to make the application behaviour self-aware.*



ANTAREX DSL is used to instrument the application for monitoring and to support autotuning.

ExaMon: Monitoring System & Power Management





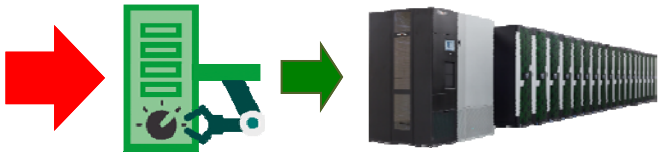
Flexible & scalable: Integrated with Linux Perf and MQTT

High-precision and low overhead (programmable events):

- 480 PMU metrics for node @ 2s : < 0.2% of overhead
- High level metrics: top-down analysis (TMAM), roofline

Suitable to be integrated in a production environment:

- Running on 64 nodes of CINECA Galileo ~ 15k metrics/s



ANTAREX

Dompé

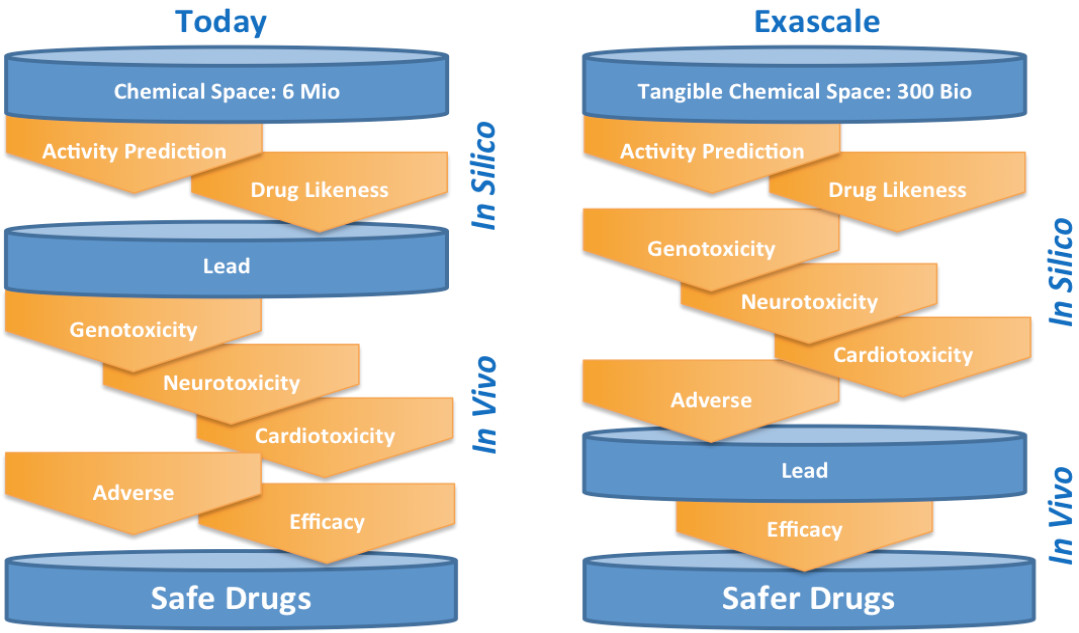
POLITECNICO MILANO 1863

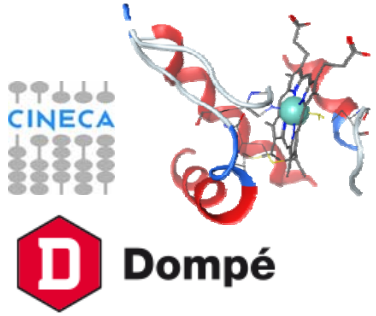
CINECA

Use Case 1: HPC Accelerated Drug Discovery System

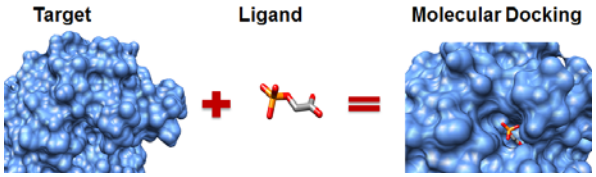
Use Case 1: HPC Accelerated Drug Discovery System

Personalized Medicine will enable to *"treat the right patient with the right drug at the right dose at the right time." [FDA]*

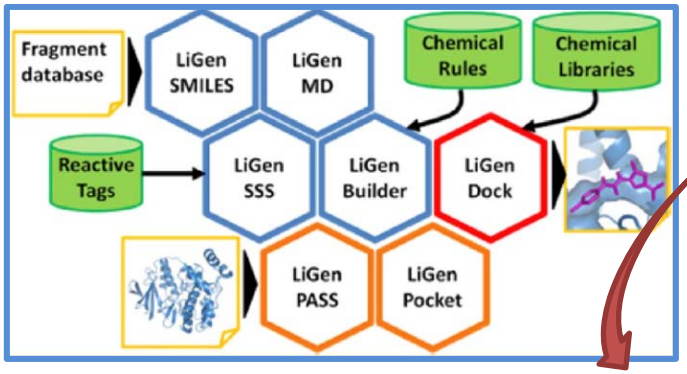




Need of HPC in Drug Discovery: HPC Molecular Simulations



LiGen in Antarex



LiGen: Exascale-ready HPC Application

ANTAREX

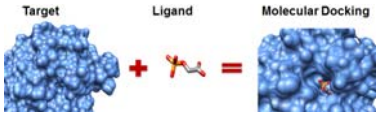
Developing **energy and resource efficient** algorithms
Using **self-functionalities to adapt and scale-out** the application

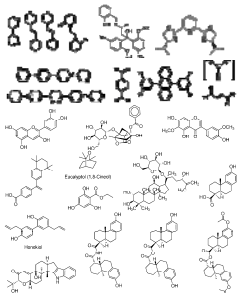
- Tangible Chemical Space: 300 Bio
- Peptide library: 120 Mio
- Commercial libraries: 6 Mio
- SafeInMan library: 9 K

Exascale HPC Virtual Screening

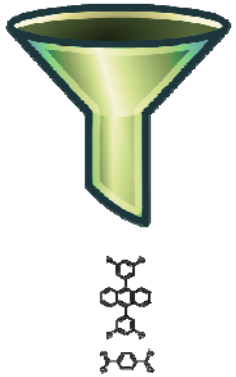
Geometric Docking

Molecular docking is a method to calculate the preferred position and shape of one molecule w.r.t a second one when bound to each other

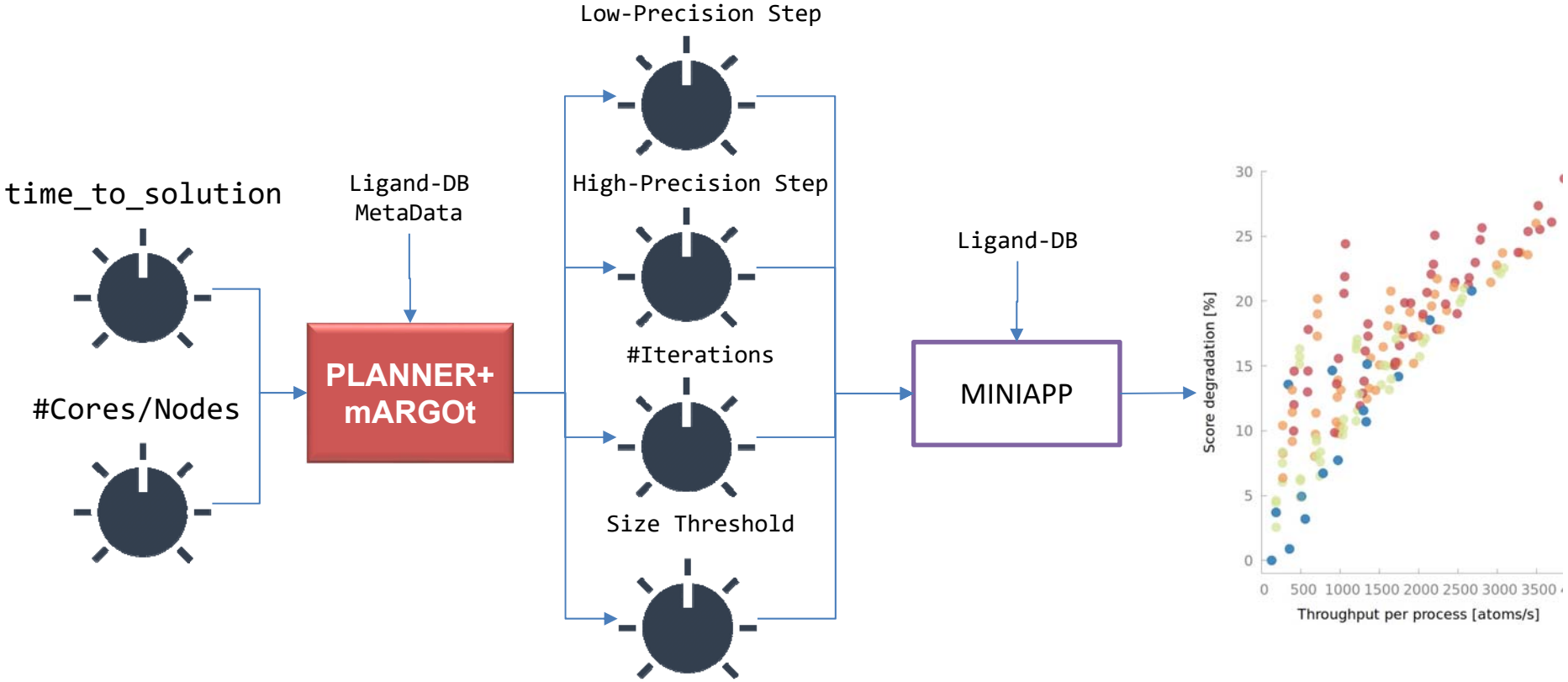


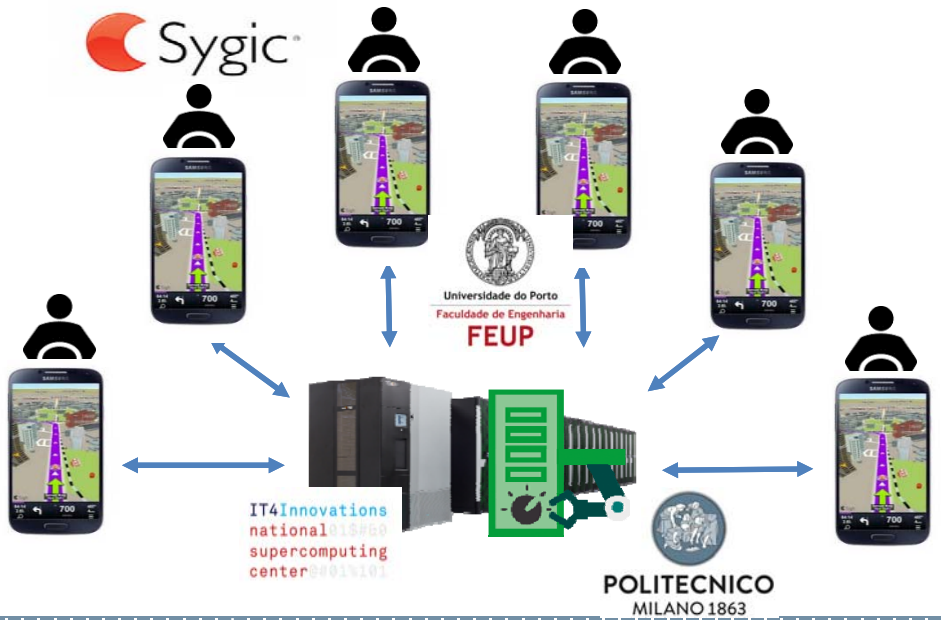


- **Geometric Docking**
  - **Shape complementarity: Geometric matching search to find out compatible pairs and most suitable poses**
- Pharmacophoric Docking
  - Molecular simulation: Exploration of a large energy landscape determined by chemical and physical interactions



Autotuning with mARGOt



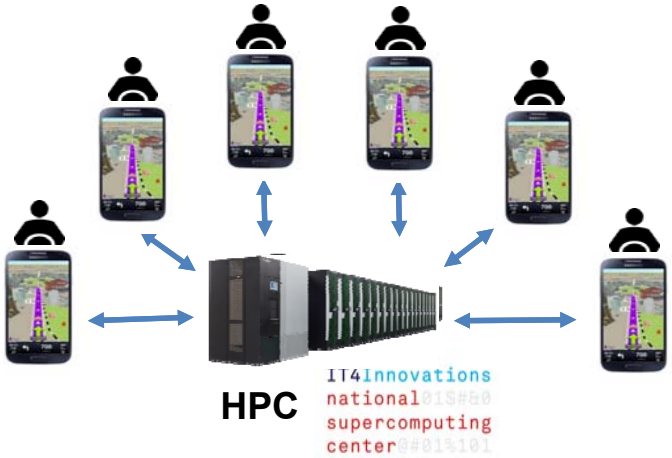


ANTAREX

Use Case 2: Self-adaptive Navigation System

Use Case 2: Self-adaptive Navigation System

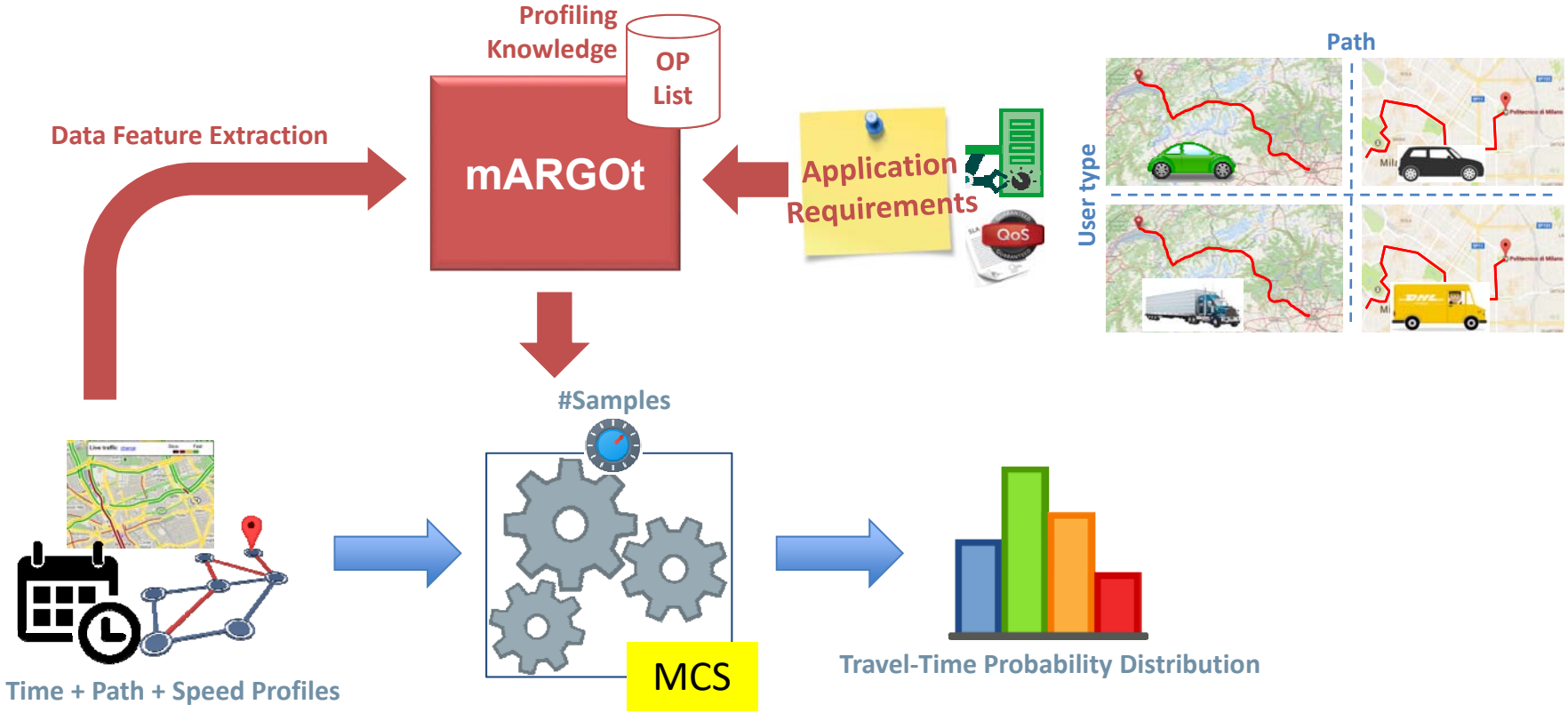
Sygic Company develops world`s most popular navigation application & provides professional navigation software for business solutions



Exploit synergies between client-side and server-side:

- Many drivers – many routing requests to HPC system
- Traffic status data sources
- Continuous update of traffic flow calculation
- ***Smart City Challenge***

Autotuning with mARGOt (User & Data-aware)



ANTAREX

European Commission | Horizon 2020 European Union funding for Research & Innovation

http://www.antarex-project.eu/

POLITECNICO MILANO 1863 | ETH zürich | Universidade do Porto Faculdade de Engenharia FEUP | Inria informatics mathematics | CINECA | IT4Innovations national supercomputing center | Dompé | Sygic

Research Collaborations