Dynamic Application Autotuning for Approximate Computing

Cristina Silvano
Multi-core and Many-core Microprocessors
What must we do differently if we are to harness them effectively?

Yale Patt
The University of Texas at Austin

Politecnico di Milano
Milano, Italia
November 13, 2008
100% computation accuracy is not always required...

Approximation offers opportunities for trading off Accuracy vs. Performance vs. Energy
The human eye stereo-matching

2 eyes → third dimension
Stereo-matching: Pixel Disparity Error vs Throughput

Left camera

Right camera

Reference Image

Stereo-matching

Application Tuning

Lower Disparity Error & Better Quality

Higher Throughput
Approximate Computing: Pareto Points

Throughput

Error

dominated design point

Pareto point

Cristina Silvano
Dynamic Autotuning of Applications

Dynamic Autotuning

At runtime according to the computation evolution

Automatic

Tuning:
1: to adjust in musical pitch or cause to be in tune: tune a guitar
2: to adjust for precise functioning: tune up an engine

In our context:
3: To adjust the values of application parameters to optimize the application metrics
Best practice is to write parametric code with **software parameters**:  
- Number of iterations  
- Application-specific parameters  
At **design-time** we extract the application knowledge:  
- Instrument the application  
- Design Space Exploration  
- Machine-learning Models  
- Store the Pareto front to get the best tradeoffs
Design-time: Learning the Application Knowledge

- Application Description
- Design of Experiments
  - Knob 1
  - Knob 2
  - Knob 3
- Design Space Exploration
- Features Clustering
  - Feature 1
  - Feature 2
  - Knobs
  - EFPs
  - Features
- Model Training and Validation
- Model Selection
  - Model not good
  - Model ok
- Generate Application Knowledge
- Broadcast Knowledge

Software-knobs domain and constraints
Extra-functional properties
Input features parameters
DoE technique and parameters
Validation parameters (MAE, R²)
Dynamic Autotuning

It is a way to constantly improve performance/energy tradeoffs with low developer effort over a wide range of run-time situations.

Dynamic SW parameters

- num sw-thread
- loop perforation
- application-specific

Cristina Silvano
It enhances a target application with an adaptation layer

- It is a C++ library to be linked to the target application
- Separation of concerns between functional and extra-functional domains.

Public repository: https://gitlab.com/margot_project/core

mARGOT Dynamic Autotuning Framework

mARGOT provides an adaptation mechanism to react to changes in:

- Application requirements
- Application-knowledge due to online learning
- System monitoring values
- Data-features extracted from input data (such as image resolution)

Public repository: https://gitlab.com/margot_project/core
What sort of society challenges could be addressed by exploiting the ANTAREX technologies?
Autotuning Geometric Docking for HPC Accelerated Drug Discovery
HPC Accelerated Drug Discovery

Need of HPC in Drug Discovery: HPC Molecular Simulations

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

LiGen HPC application for drug discovery

Cristina Silvano
**Molecular docking** is a method to estimate the preferred 3D position and shape of a candidate drug (ligand) in the target protein pocket when bound to each other

- **Geometric Docking**
  - **Shape Complementarity**: 3D geometric matching search to find out compatible pairs and most suitable poses

- **Pharmacophoric Docking**
  - **Molecular Simulation**: exploration of a large energy landscape given by chemical & physical interactions

Cristina Silvano
Expose software-parameters from the geometric docking application
Marconi: the most powerful public supercomputer in Italy

- **No. 19 in Top500 and No. 4 in Europe:** Marconi Intel Xeon Phi: 10.38 PetaFlops (Linpack performance) 18.8 PetaFlops (peak performance) with 348,000 cores. Site: Casalecchio di Reno, Bologna (Italy)

- **Marconi** is the Cineca's Tier-0 system, co-designed by Cineca and Lenovo based on the Lenovo NeXtScale platform and Intel® Xeon Phi™ product family alongside with Intel® Xeon® processor and Intel Omni-Path.
LiGen GeoDock: The Zika virus experiment

EXSCALATE: ExaSCale smArt pLatform Against pathogEns

- 1.2 Billion ligands dataset (candidate drugs)
- 26 Zika binding sites (pharmacological targets)
- 8 Trillion poses scored (by GeoDock)
- 260 TeraByte of stored data
- About 900K Threads on 300k cores on 10 petaFLOPs MARCONI
- 1 MW measured power consumption
- Run Time to Solution: 3.2 h for 1 out of 26 sites (run in Jan 2019)
- Total Time to Solution: 3.5 days (84 h) for 26 sites
- Energy to Solution: 84 MWh

Estimated Exascale Run in 2021: from 84 h to less than 1 h

Experiment website:
https://www.antarex4zika.eu
Autotunining an HPC-based Navigation System for Smart Cities
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**

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

**Sygic Top #2 App in navigation category worldwide with 200 M users**

**Sygic world’s 1st for iPhone, 2nd for Android**
Motivations:
• Provide optimal routes to hundreds of thousands of drivers/cars operating in the city area
• Serve all drivers’ requests with global best to reduce total driving time
• Avoid traffic jams

Requires:
• Intelligent routing based on accurate calculation of traffic view state
• Balance routes for a city global optimum
• Minimize data transfer

Intelligent Navigation for Smart Cities
Cristina Silvano
What is the Probabilistic Time-dependent Routing for a navigation system?

- Module to evaluate the expected travel time
- In a client-server navigation infrastructure, the server-side must evaluate accurate expected travel time with updated traffic information
- Implemented by a MonteCarlo Simulation (MCS) to evaluate the probabilistic speed profile for each hop
- Dynamic autotuning of the number of samples for the MCS

-European Innovation Radar Prize 2018-

- Best Early Stage Innovation

- **ICT 2018 Vienna Award**
  - Top 20 Innovations in 2018

  “Saves money to drivers and cities. Contributes to reduction of CO2 emissions. Improves quality of life in urban areas. Reduces time spent in daily travel traffic by more than 20 percent”