Computing on Low-Power Architectures (COLA)

Report dei Contributi

https://agenda.infn.it/e/10434

ID contributo: **0** Tipo: **non specificato**

Introduction

Autore principale: MANTOVANI, Filippo (BSC)

ID contributo: **1** Tipo: **non specificato**

Welcome and Opening

giovedì 25 febbraio 2016 08:45 (15 minuti)

Relatore: PARESCHI, Lorenzo (University of Ferrara)

ID contributo: **2** Tipo: **non specificato**

ARM in HPC: Software and Tools

giovedì 25 febbraio 2016 09:30 (25 minuti)

I will briefly update the audience on the state and availability of ARM tools for HPC, covering compilers, debuggers and profilers, and briefly summarise hardware availability.
What then follows is a presentation of what I perceive to be the main issues that ARM needs to from a software tools perspective in HPC. Some of these are specific to ARM, and some are industry-wide issues, where ARM could potentially differentiate itself by leading the way to an innovative solution.

Relatore: NORTH, Geraint (ARM)

ID contributo: **3** Tipo: **non specificato**

Computing on Low-Power Architectures

giovedì 25 febbraio 2016 09:00 (30 minuti)

Relatore: MANTOVANI, Filippo (BSC)

ID contributo: **5** Tipo: **non specificato**

Porting HPC Libraries and Applications to ARM's 64-bit Architecture

giovedì 25 febbraio 2016 10:00 (25 minuti)

Last year at SuperComputing 2015 ARM announced the ARM Performance Libraries providing BLAS, LAPACK and FFT routines optimised for the ARM AArch64 architecture. Alongside this announcement ARM gave out a list of open source HPC libraries and applications that it would be shipping. This talk will go over some of the issues faced along the road to porting and providing all of these packages.

Relatore: LANDER, George (ARM)

ID contributo: **6** Tipo: **non specificato**

The INFN COSA project

giovedì 25 febbraio 2016 11:00 (25 minuti)

The embedded and high-performance computing sectors have in the past been very isolated and unaware of each other's needs and technologies. Similar isolations have occurred between HPC and the mobile/tablets commodity markets. We are now experiencing a very important convergence between markets, both in constraints and needs as well as in technologies. High computational demands, power consumption limitation, parallelism, heterogeneous computing and cost effectiveness are now driving constraints of both the HPC and embedded sectors. This convergence opens the way to the possibility of performing scientific computation on low power architecture originally developed for the embedded or mobile world. In this talk, we present the panorama of the low power architectures suitable for scientific computation. The INFN experience in building a low power cluster based on System-on-Chips (SoCs) is discussed together with the performance results in terms of power ratio and energy consumption obtained on that cluster. The applications used in the tests range from synthetic benchmarks to real life use cases. Results are compared to those obtained on traditional HPC architectures.

Relatore: CESINI, Daniele (INFN - CNAF)

ID contributo: **7** Tipo: **non specificato**

Energy to Solution vs Time to Solution, towards energy-aware HPC applications

giovedì 25 febbraio 2016 11:30 (25 minuti)

Energy efficiency is quickly gaining importance in the HPC field.
High-end processors are evolving towards more advanced power-saving and power-monitoring technologies, while low-power processors, designed for the mobile market, are gaining interest in the HPC area thanks to their increasing computing capabilities, in conjunction with their competitive pricing.
On the other hand, from the software point of view, HPC applications are still optimized mainly for performance, often neglecting energy considerations, despite the fact that data-centers in the near future may start to account for consumed energy, instead of running time.
In this work we explore how HPC applications may became more energy-aware; in particular we explore energy-profile benchmarks of actual HPC applications on different architectures, in order to compare their energy performance, but also to identify different available software strategies to tune energy consumption.

Relatore: CALORE, Enrico (UNIFE and INFN)

ID contributo: **8** Tipo: **non specificato**

Low Power processor in HEP

giovedì 25 febbraio 2016 12:00 (25 minuti)

High Energy Physics benefits from an implicit parallelism at the level of the single physics event. Each event can be processed indipedentely making very easy the distribution of the event on a cluster of independent computing node. The problem is the huge number of events that requires thousand of power hungry worker node. The HEP community is starting to look at even bigger number of smaller but energy efficient processors. The talk will concentrate on the reference benchmark for HEP, called HS06 and the relative performance of present processor in term of HS06/Watt.

Relatore: MICHELOTTO, Michele (INFN)

ID contributo: **9** Tipo: **non specificato**

Low Power Computing in Gamma-Ray Astronomy

giovedì 25 febbraio 2016 12:30 (25 minuti)

Gamma-Ray Astronomy is an optimal test-ground for Low-Power Computing and High-Throughput Computing. On the one hand, ground based detectors for Gamma-ray Astronomy are the prototypes for distributed experiments, as single detectors may be scattered in an area of few square kilometres, and the capability of each unit to process, at least partially, its own data before sending them to the central data acquisition provides a key advantage. On the other hand, satellite-born detector needs low-power on-board and huge computing power facilities for the ground processing. The talk will present some applications in the field of Gamma-Ray Astronomy, ranging from a GPU chain to build the model that best represent the data acquired in space (by evaluating the Maximum Likelihood Ratio), to an FPGA/ARM architecture used to process images collected by Cherenkov telescopes on ground.

Relatore: BASTIERI, Denis (UNIPD and INFN)

ID contributo: **10** Tipo: **non specificato**

Near-threshold Scalable Computing - The PULP experience

giovedì 25 febbraio 2016 14:00 (25 minuti)

The “internet of everything”envisions trillions of connected objects loaded with high-bandwidth sensors requiring massive amounts of local signal processing, fusion, pattern extraction and classification. Higher level intelligence, requiring local storage and complex search and matching algorithms, will come next. From the computational viewpoint, the challenge is formidable and can be addressed only by pushing computing fabrics toward massive parallelism and brain-like energy efficiency levels. We believe that CMOS technology can still take us a long way toward this vision. Our recent results with the PULP (parallel ultra-low power) open computing platform demonstrate that pj/OP (GOPS/mW) computational efficiency is within reach in today’s 28nm CMOS FDSOI technology. In the longer term, looking toward the next 1000x of energy efficiency improvement, we will need to fully exploit the flexibility of heterogeneous 3D integration, stop being religious about analog vs. digital, Von Neumann vs. “new”computing paradigms, and seriously look into relaxing traditional “hardware-software contracts”such as numerical precision and error-free permanent storage.

Relatore: BENINI, Luca (ETHZ and UNIBO)

ID contributo: **11** Tipo: **non specificato**

Energy-Aware Scheduling at the Leibniz Supercomputing Centre

giovedì 25 febbraio 2016 14:30 (25 minuti)

Due to rising energy prices and increasing carbon footprint, it is commonly accepted that the main constraint for future, sustainable many-Peta to Exascale HPC system will be dictated by power consumption. Along with the design of more energy-efficient hardware and cooling infrastructures, a viable way of addressing this challenge is offered by energy-aware scheduling. This presentation explains the approach adopted by the Leibniz Supercomputing Centre to reduce power consumption by employing energy aware management software and thorough power consumption monitoring. Specifically, we will describe the energy aware scheduling feature of IBM LoadLeveler, the resource and management system adopted in SuperMUC, one of the faster supercomputers in the world. This feature allows to select the most "energy-efficient" CPU frequency for a large fraction of SuperMUC's application portfolio and, therefore, contributes to substantially reducing the overall energy consumption of the system.

Relatore: TAFANI, Daniele (Leibniz Supercomputing Center)

ID contributo: **12** Tipo: **non specificato**

Climbing Mont Blanc - a Prototype System for Training in Energy Efficient Programming

giovedì 25 febbraio 2016 15:00 (25 minuti)

Climbing Mont Blanc (CMB) is an open online judge used for training in energy efficient programming of state-of-the-art heterogeneous multicores. It is based on an Odroid-XU3 board with an Exynos Octa processor and integrated power sensors. The system currently accepts C and C++ programs, with support for OpenCL v1.1, OpenMP 4.0 and Pthreads. Programs submitted using the graphical user interface are evaluated with respect to performance and energy-efficiency. A small and varied set of problems are available. We are not aware of any other online programming judges that reports energy-efficiency. The talk will present some early experience from using the CMB system, potential opportunities for collaboration and future work. Our long term goal is to learn more about energy-efficient computing on handheld devices from submissions to the system.

Relatore: NATVIG, Lasse (Norwegian University of Science and Technology)

ID contributo: **13** Tipo: **non specificato**

Execution of the DPSNN spiking neural network simulator on the nVIDIA Jetson TK1 platform

giovedì 25 febbraio 2016 15:30 (25 minuti)

Fast simulation of spiking neural network models plays a dual role: it contributes to the solution of a scientific grand-challenge –i.e. the comprehension of brain activity –and, by including it into embedded systems, it can enhance applications like autonomous navigation, surveillance and robotics.
The DPSNN is a spiking neural network simulator developed at the INFN APE lab. It is coded as a network of C++ processes, and it is designed to generate spiking behaviors and synaptic connectivity that do not change when the number of processing nodes is varied, easing the quantitative study of scalability.
We used the DPSNN as a benchmark for the ARM-based nVIDIA Jetson TK1 platform measuring instantaneous power, total energy consumption, execution time and energetic cost per synaptic event.
Results will be presented and compared against those obtained on an Intel Xeon platform.

Relatore: LONARDO, Alessandro (INFN)

ID contributo: **14** Tipo: **non specificato**

Quantum ESPRESSO community code and the Exascale Challenge

giovedì 25 febbraio 2016 16:30 (25 minuti)

QUANTUM ESPRESSO builds upon electronic-structure codes that have been developed and tested by some of the original authors of novel electronic-structure algorithms and applied in the last twenty years by some of the leading materials modeling groups worldwide.
Innovation and efficiency are its main focus, with special attention paid to massively parallel architectures, and in the exascale challenge has been selected by many HPC centers, and technology providers world-wide as one of the application worth to be ported on new architecture. In the talk the refactoring effort and the porting strategies, toward exascale, will be presented and discussed together with preliminary results on new highly parallel chips.

Relatore: CAVAZZONI, Carlo (CINECA)

ID contributo: **15** Tipo: **non specificato**

High throughput data acquisition with InfiniBand on low power architectures

giovedì 25 febbraio 2016 17:00 (25 minuti)

LHCb experiment is preparing a major upgrade, during long shutdown 2 in 2018, of both the detector and the data acquisition system. A system composed of about 500 nodes and capable of transporting up to 50 Tbps of data will be required, this can only be achieved in a manageable way using a readout system based on commodity hardware and high-bandwidth data-centre switches. Several studies are ongoing in order to investigate different network and hardware technologies with the aim of reducing the purchase and maintenance costs of the system. In this presentation we will introduce InfiniBand and show preliminary tests with this network technology and x86 low power architectures. We will also describe how optimisations, like the usage of core-affinity, can affect the performances of such kind of systems.

Relatore: MANZALI, Matteo (UNIFE)

ID contributo: **16** Tipo: **non specificato**

Wrap Up Day1

giovedì 25 febbraio 2016 18:00 (30 minuti)

ID contributo: **17** Tipo: **non specificato**

Squeezing Deep Learning onto a Phone

venerdì 26 febbraio 2016 12:00 (25 minuti)

Deep learning has recently emerged as one of the most promising techniques for classification, with breakthrough results in fields such as image recognition and natural language processing. However, deep learning calls for a tremendous amount of resources, chiefly in the training phase, but also during the inference phase. This may not be an issue when "Google-scale" computing facilities are available, but it hampers the applicability of deep learning in several fields where computing power, or memory capacity, or energy, are constrained. The talk will focus on one such field: mobile computing, where it is clear that client-side machine learning algorithms will play a key role in the next generation of applications, but nontrivial progress is required to tailor such algorithms to a limited computing or power budget. The talk will present preliminary observations and measurements taken from an image recognition application with convolutional neural networks. As HPC is also being infiltrated by mobile technologies, with ARM processors expected to appear in the Green500 list any time soon, such observations acquire a more general significance.

Relatore: FANTOZZI, Carlo (UNIPD)

ID contributo: **18** Tipo: **non specificato**

Structured Parallel Programming on multi-core wireless sensor networks

venerdì 26 febbraio 2016 09:30 (25 minuti)

Wireless sensor network (WSN) platforms are now experiencing the same evolution of high performance computing (HPC) when it evolved from singlecore to multi-core architectures. Multi-core sensor platforms are expected to grow, especially in application domains that require complex processing of the sensed data, such as those that require image processing, data encryption, network coding, data fusion etc.

The shift from single-core to multi-core sensor platforms also affects the WSN programming models. In fact, it introduces the need of high level abstractions to support parallel and distributed programming and models in WSN. This fact has recently suggested the adoption in WSN of methodologies such as skeletons that are largely used in the programming of parallel and distributed systems.

Our work addresses the use of skeletons in the context of WSN, with the particular attention to multi-core sensors. In particular, leveraging on the fact that some meaningful WSN applications are characterised by known programming patterns (for example, in visual sensor networks, the stencil skeleton fits well object tracking applications), we aim at defining suitable models of computation for the most promising skeletons for WSN, and at combining the concepts of structured parallel programming and real-time sensing.

Relatore: CHESSA, Stefano (UNIPI)

ID contributo: **19** Tipo: **non specificato**

Porting and testing the Einstein Toolkit on the the generation of low-power architectures.

venerdì 26 febbraio 2016 10:00 (25 minuti)

Low-Power architectures are subject of much interest also as viable alternatives to traditional HPC platform. In this talk we will focus on
the performance that can now be obtained porting a large simulation toolkit (The EinsteinToolkit), widely used in Numerical Astrophysics
to simulated matter coupled to the Einstein's equations, to Low Power Architectures. We considered multicores / multi node cluster based on ARM and Intel low power processors and we compared results with a traditional HPC cluster, the Galileo system at CINECA. The work has been performed using the resources actually available for the INFN-COSA project.

Relatore: DE PIETRI, Roberto (UNIPR and INFN)

ID contributo: **20** Tipo: **non specificato**

Experience with Beignet OpenCL on low power Intel SoC

venerdì 26 febbraio 2016 11:00 (25 minuti)

This presentation will focus on our first-hand experience in running benchmarks using Open Source OpenCL, Beignet, on both the GPU and CPU of a low power Intel Skylake SoC.

Relatore: PANTALEO, Felice (CERN)

ID contributo: **21** Tipo: **non specificato**

Experience running codes on ARM64+GPU platforms

venerdì 26 febbraio 2016 11:30 (25 minuti)

The presentation is going to be focused on the first-hand experience in running CUDA-accelerated applications on ARM64 platforms with NVIDIA GPU Kepler cards. The talk will underline challenges, difficulties, weakness and strength of an heterogeneous platform.

Relatore: SPIGA, Filippo

ID contributo: **22** Tipo: **non specificato**

GPU programming for complex fluids

venerdì 26 febbraio 2016 09:00 (25 minuti)

In this contribution we will discuss issues related to the optimisation of Lattice Boltzmann multi-component flow solver to study the physics of soft glassy system on multi-GPU platforms.

Relatore: KUMAR, Pinaki (Technische Universiteit Eindhoven)

ID contributo: **23** Tipo: **non specificato**

Wrap Up Day2

venerdì 26 febbraio 2016 12:30 (30 minuti)

ID contributo: **24** Tipo: **non specificato**

Closed session (TPC meeting, …)

venerdì 26 febbraio 2016 14:00 (4 ore)

ID contributo: **25** Tipo: **non specificato**

Exploration of Future Computing Platforms at CMS

giovedì 25 febbraio 2016 17:30 (25 minuti)

Overview of various efforts at Compact Muon Solenoid (CMS) experiment at CERN on emerging general-purpose computing platforms for High Throughput Computing (HTC). We report our experience on software porting, performance, energy efficiency and building a demonstrator Worldwide LHC Computing Grid (WLCG) Tier-3 computing site at Princeton University based on ARMv8 64-bit Server-on-Chip.

Relatore: ABDURACHMANOV, David (CERN)