Journée Architecture / Compilation

Jeudi 1er octobre 2015

Amphitéatre 33, CEA NanoInnov,

Av. de la Vauve, 91120 Palaiseau

10:30-11:30 : Benoit Dupont de Dinechin
Séminaire : « Revisiting DSP Acceleration with the Kalray MPPA Manycore Processor »

14:00-16:00 : Alexandre Aminot
Soutenance de thèse : « Placement de tâches dynamique et flexible sur processeur multi coeur asymétrique en fonctionnalités»

10:30 Présentation invitée
Titre : « **Revisiting DSP Acceleration with the Kalray MPPA Manycore Processor »**

Abstract : General-purpose CPU acceleration is motivated by increased performance or energy efficiency in the areas of computing, networking and storage. The mainstream acceleration technologies are based on FPGA, DSP, GPU and manycore architectures such as Intel MIC. FPGA and DSP accelerators are traditionally applied to embedded computing applications, or to network and storage processing functions. GPU and manycore accelerators are mostly deployed in high-performance computing systems and datacenters. While the energy efficiency benefits of those different classes of accelerators appears comparable, FPGA and DSP accelerators are distinguished by predictable execution times and low-latency I/O, while GPU and manycore accelerators offer the highest floating-point performances.

We present the architecture, software environment and target application areas of the MPPA-256 Bostan processor, the second generation implementation of the Kalray manycore architecture. The key challenges were to overcome the limitations of classic DSP acceleration, without losing the distinguishing advantage of a precision-timed computing machine. The MPPA-256 Bostan processor operates between 400MHz and 800MHz, for a typical power consumption between 10W and 20W. Its peak floating-point performances at 600MHz are 634GFLOPS and 316GFLOPS for single- and double-precision respectively. Two DDR3 memory controllers at 2133 MT/s provide an aggregate external memory bandwidth of 34GB/s. The processor also provides two PCIe Gen3 8-lane interfaces, 8 Ethernet 10Gb/s links and a direct access to the network-on-chip (NoC) for low-latency processing of data streams. The MPPA-256 Bostan processor architecture is clustered with 16 compute clusters and 2 I/O clusters, where each cluster is built around a multi-banked local static

memory (SMEM) shared by the 16+1 (compute cluster) or 4+4 (I/O cluster) processing (PE) + management (RM) cores. Only the quad-cores in the I/O clusters can directly address the external DDR memory. This memory hierarchy is energy-efficient and prevents timing interferences between the clusters, except for the explicit data moving through the NoC.

At the cluster level, the software environment is based on the GNU C/C++ compilers with OpenMP and the associated language libraries. Linux SMP is running in the I/O tiles along with a POSIX RTOS in the compute tiles; however these operating systems are para-virtualized on top of an exo-kernel type of hypervisor. This hypervisor manages the NoC and provides the foundations of a software distributed shared memory (DSM) run-time system, which leverages the MMU of each core to enable direct addressing into the external DDR memory. This software DSM effectively turns the local memory of each compute tile into a last-level cache. It also enables one or more MPPA-256 Bostan processors connected by their NoC extensions to appear as a single OpenCL device to a host CPU, with the global memory implemented by stripping data across the DDR controllers.

14:00 Soutenance de thèse
Titre : **« Placement de tâches dynamique et flexible sur processeur multi coeur asymétrique en fonctionnalités »**

Abstract : Pour permettre une meilleure scalabilité des architectures *manycoeur*, cette thèse s'intéresse aux architectures asymétriques en fonctionnalités. Cette asymétrie est réalisée par la distribution non-uniforme des extensions matérielles (ex. FPU, SIMD) à travers les cœurs. Les avantages en surface sont apparents, mais qu'en est-il de l'impact au niveau énergétique et performance ? Pour répondre à cette question, la thèse explore la nature de l'utilisation des extensions dans des applications de l'état de l'art et les différentes méthodes existantes. Pour optimiser le placement de tâches et ainsi augmenter l'efficacité, la thèse propose une solution dynamique au niveau ordonnanceur.

Devant le jury composé de :

- **Guy GOGNIAT** Professeur Université de Bretagne-Sud - UEB Rapporteur
- **Alain PEGATOQUET** Maître de conférences Univ. de Nice Sophia Antipolis Rapporteur
- **Albert COHEN** Directeur de recherche INRIA - ENS Examinateur
- **Karine HEYDEMANN** Maître de conférences Univ. Pierre et Marie Curie Examinateur
- **Vivien QUEMA** Professeur Grenoble INP / ENSIMAG Examinateur
- **Benoit Dupont de DINECHIN** Directeur de la technologie Kalray Examinateur
- **Henri-Pierre CHARLES** Directeur Recherche CEA LIST Directeur de thèse
- **Yves LHUILLIE**R Ingénieur Recherche CEA List Encadrant de thèse
- **Andrea CASTAGNETI** Ingénieur Recherche CEA List Co-Encadrant de thèse (invité)