HAL
open science

X-Kaapi: a Multi Paradigm Runtime for Multicore Architectures

Thierry Gautier, Fabien Lementec, Vincent Faucher, Bruno Raffin

To cite this version:

Thierry Gautier, Fabien Lementec, Vincent Faucher, Bruno Raffin. X-Kaapi: a Multi Paradigm Runtime for Multicore Architectures. [Research Report] RR-8058, 2012, pp.16. hal-00727827v1

HAL Id: hal-00727827

https://inria.hal.science/hal-00727827v1

Submitted on 4 Sep 2012 (v1), last revised 17 Dec 2013 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

X-KAAPI: a Multi Paradigm Runtime for Multicore Architectures

Thierry Gautier, Fabien Lementec, Vincent Faucher, Bruno Raffin

RESEARCH REPORT
N° 8058
Feb 2012
Project-Teams MOAIS

ISSN 0249-6399 ISRN INRIA/RR--8058--FR+ENG

X-Kaapi: a Multi Paradigm Runtime for Multicore Architectures

Thierry Gautier*, Fabien Lementec†, Vincent Faucher‡, Bruno Raffin§

Project-Teams MOAIS

Research Report n° 8058 — Feb 2012 — 16 pages

Abstract: The paper presents X-Kaapi, a compact runtime for multicore architectures that brings multi parallel paradigms (parallel independent loops, fork-join tasks and dataflow tasks) in a unified framework without performance penalty. Comparisons on independent loops with OpenMP and on dense linear algebra with QUARK/PLASMA confirm our design decisions. Applied to EUROPLEXUS, an industrial simulation code for fast transient dynamics, we show that X-Kaapi achieves high speedups on multicore architectures by efficiently parallelizing both independent loops and dataflow tasks.

Key-words: parallel computing, X-Kaapi, Europlexus

* INRIA
† INRIA
‡ CEA, DEN, DANS, DM2S, SEMT, DYN, F-91191 Gif-sur-Yvette, France
§ INRIA

RESEARCH CENTRE GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

X-Kaapi: un support exécutif multi-paradigme pour architecture multi-cœur

Résumé : Ce rapport présente X-Kaapi, un support exécutif pour architecture multi-cœur qui permet l'exploitation conjointe de plusieurs paradigmes de programmation parallèle (boucles indépendantes, fork-join, flot de données). Les surcoûts à l'exécution sont faibles et nous présentons des comparaisons pour la programmation de boucles indépendantes avec OpenMP, et sur des problèmes en algèbre linéaire dense nous nous comparons à QUARK/-PLASMA. Enfin nous présentons les résultats obtenus lors de la parallélisation du code EUROPLEXUS de dynamique rapide et qui utilise plusieurs de ces paradigmes.

Mots-clés : environment de programmation parallèle, X-Kaapi, Europlexus

1 Introduction

Industrial codes usually require mixing different parallelization paradigms to achieve interesting speedups. The challenge is to develop programming and runtime environments that efficiently support this multiplicity of paradigms. We introduce X-KAAPI, a runtime for multicore architectures designed to support multiple parallelization paradigms with high performance thanks to a low overhead scheduler. Our case study is the industrial numerical simulation code for fast transient dynamics called EUROPLEXUS. EUROPLEXUS[1,2] is dedicated to complex simulations in industrial framework, with a large source code composed of 600.000 lines of Fortran. It supports 1-D, 2-D and 3-D models, based on either continuous or discrete approaches, to simulate structures and fluids in interaction. EUROPLEXUS supports non-linear physics for both geometrical (finite displacements, rotations and strains) and material (plasticity, damage, etc) properties. A typical simulation spends more than 70% of the execution time in:

1. large loops with independent iterations,
2. Sparse Cholesky matrix factorizations in a Skyline storage format.

As EUROPLEXUS is mainly used to simulate impacts, structures are subject to important deformations, leading to changing and unbalanced work loads.

Reaching high performance on multicore architectures requires several threads of control running mostly independent code, with few synchronizations to ensure a correct progress of the computation. Programming directly with threads is highly unproductive and error prone [16]. Two main programming alternatives have been developed. Cilk [3] promotes a fork-join parallel paradigm with theoretical guarantees on the expected performance. OpenMP [17] relies on code annotations to generate parallel programs. Cilk and OpenMP have both basic constructs to parallelize independent loops. With OpenMP the user has also the ability to guide the way iterations are scheduled among the threads.

Ten years after Cilk, the introduction of tasks in OpenMP-3.0 makes Cilk and OpenMP, at a first glance, very close. They seem to be good candidates for a task based Cholesky factorization [13, 1], but the current implementation of tasks in OpenMP-3.0 [17] (in Intel compiler or GCC compiler) is several orders of magnitude more costly than in Cilk, making it hard to reach portable performance because of the grain size decision problem [3, 6].

Moreover, [13, 6] show that OpenMP-3.0 and Cilk parallel models limit the available parallelism for a dense Cholesky factorization. The authors promote a data flow runtime that is able to encode finer data flow synchronizations between tasks. The runtime can detect concurrent tasks as soon as

[1] http://europlexus.jrc.ec.europa.eu/public/manual_html/index.html
[2] http://www.repdyn.fr

their inputs are produced. Such data flow programming model is a promising approach for our sparse Cholesky factorization.

Several runtimes and languages were based on a data flow paradigm, like Athapascan [9] used for sparse Cholesky factorizations [5], QUARK [22], the data flow runtime of the PLASMA dense linear algebra library [4], the StarSS programming model with its SMP implementation called SMPSs [2], or StarPU [1] dedicated to multi-GPU computations. But none of these softwares support independent loops. Moreover, they do not allow the creation of recursive tasks, discarding recursive parallel algorithms.

The X-Kaapi runtime we introduce in this paper proposes a new unified framework based on data flow tasks and workstealing dynamic scheduling to develop multi-paradigms fine grain parallel programs. A comparison with OpenMP shows that our dynamic scheduler can outperform both the static and dynamic OpenMP scheduler. We also used X-Kaapi to develop a binary compatible QUARK library to schedule PLASMA's algorithms with better scalability at a finer grain. Finally, we report preliminary results mixing both parallel loops and data flow task parallelism in EUROPLEXUS.

Next section presents the X-Kaapi's parallel programming model. We focus on its adaptive task model, and how it is used to support parallel loops. Section 3 reports experimental evaluations compared to OpenMP [17] on parallel loops and QUARK [22] on dense Cholesky factorizations. Section 4 evaluates the parallelization of EUROPLEXUS as compared with OpenMP, before concluding.

2 Data flow task programming with X-Kaapi

The X-Kaapi's task model [10] , as for Cilk [3], Intel TBB [19], OpenMP-3.0 [17] or StarSs/SMPSs [2], enables non blocking task creation: the caller creates the task and continues the program execution. The semantic remains sequential like for its predecessor Athapascan [9], but the runtime was redesigned [10] and the task model extended to support adaptive tasks (section 2.4).

The execution of a X-Kaapi program generates a sequence of tasks that access to data in a shared memory. From this sequence, the runtime extracts independent tasks to dispatch them to idle cores. We focus here on the multicore version of X-Kaapi.

2.1 Design choices

More than a runtime, X-Kaapi[3] is a fully featured software stack to program heterogeneous parallel architectures. The stack is written in C and was designed using a bottom up approach: each layer is kept as specialized as

[3] `http://kaapi.gforge.inria.fr`

possible to fit a specific need. Currently, the stack includes: a runtime supporting multicores and multiprocessors; a set of ABIs (QUARK [22], OpenMP runtime libGOMP); a set of high level APIs (C [14], Fortran and C++; subset of Intel TBB [19]); and a source to source compiler [15] based on the ROSE framework [18].

2.2 Data flow task model

A X-Kaapi program is composed of sequential C or C++ code and some annotations or runtime calls to create tasks. The parallelism in X-Kaapi is explicit, while the detection of synchronizations is implicit [10]: the dependencies between tasks and the memory transfers are automatically managed by the runtime.

A task is a function call that returns no value except through the shared memory and the list of its effective parameters. Depending of the APIs, tasks are created using code annotation (`#pragma kaapi task` directive) if the X-Kaapi's compiler [15] is used, or by library function (`kaapic_spawn` call using X-Kaapi's C API [14]), or by low level runtime function calls.

Tasks share data if they have access to the same memory region. A memory region is defined as a set of addresses in the process virtual address space. This set has the shape of a multi-dimensional array. The user is responsible for indicating the mode each task uses to access memory: the main access modes are *read*, *write*, *reduction* or *exclusive* [9, 10, 15, 14]. When required [10], the runtime computes true dependencies (Read after Write dependencies) between tasks thanks to the access modes. At the expense of memory copy, the scheduler may solve false dependencies through variable renaming.

A thread creates tasks and pushes them in its own workqueue. The workqueue is represented as a stack. The enqueue operation is very fast, typically about ten cycles on the last x86/64 processors. As for Cilk, a running X-Kaapi's task can create child tasks, which is not the case for the other data flow programming softwares previously mentioned [22, 2, 1]. Once a task ends, the runtime executes the children following a FIFO order. During task execution, if a thread encounters a stolen task, it suspends its execution and switches to the workstealing scheduler that waits for dependencies to be met before resuming the task. Otherwise, and because sequential execution is a valid order of execution [9, 10], tasks are performed in FIFO order without computation of data flow dependencies.

2.3 Execution with workstealing algorithm

X-Kaapi relies on *workstealing*, popularized by Cilk [3], to dynamically balance the work load among cores. Once a thread becomes idle, it becomes a thief and initiates a steal request to a randomly selected victim. On reply,

the thief receives one or more ready tasks. X-KAAPI favors *request aggregation* [11]: N pending requests to a same victim are handled in one operation, reducing the number of ready task detections. A theoretical analysis in [20] shows a reduction of the total steal request number. In our protocol, one of the thieves is elected to reply to all requests.

As opposed to Cilk, X-KAAPI considers tasks with data flow dependencies. Following the *work first principle* [8], X-KAAPI computes ready tasks on steals, favoring work at the expense of the critical path. The detection of a ready task consists in a traversal of the victim stack from the top most task (the oldest), to look all its predecessors have been completed. Following the Cilk's T.H.E protocol [8], X-KAAPI synchronizes the thief and victim using a Dijkstra' protocol. Except in rare cases, the victim and the thief execute concurrently. Using this approach, X-KAAPI and Cilk show similar overheads for the execution of independent tasks (see section 3.1).

The overhead to manage tasks and to compute the data flow graph may remain important. To reduce this overhead, X-KAAPI implements two original optimizations.

First, when the cost of computing ready tasks becomes important, the runtime attaches to the victim an accelerating data structure for steal operations. The structure contains a list that gets updated with tasks becoming ready due to the completion of their data flow dependencies. A subsequent steal operation is reduced to the pop of a task from the ready list (nearly constant time operation), without a traversal of the victim stack.

The second optimization enables a more fundamental reduction of parallelism overhead. Parallel versions of some algorithms require more operations than their sequential counterpart. The overhead is directly related to the number of created tasks. The idea is thus to limit the number of tasks by creating them on demand, as computing resources become idle. These so called *adaptive* tasks are detailed in the following section.

2.4 Adaptive task model

Writing performance-portable programs within the task programming model requires creating much more tasks than available computing resources. Then, the scheduler can efficiently and dynamically balance the work load. However, the extra operations required to merge the partial results account for overhead since it is not present in the sequential algorithm. Fich [7] proved that any parallel algorithm of time $\log n$ to compute prefix of n inputs requires at least $4n$ operations, versus $n-1$ operations in sequential. Adapting the number of created parallel tasks to dynamically fit the number of available resources is the key point to reach high performance. With an other approach for implementing this adaptation, we have proposed this on demand task creation to build coarse grain parallel adaptive algorithms for most of the STL algorithms [21]. Here, the proposed solution extends the

task model for a much more finer integration with the scheduler.

In data flow model, once all inputs of a task are produced, it becomes ready for execution. A task being executed cannot be stolen. To allow on demand task creation, X-KAAPI extends this model: a task publishes a function, called the *splitter*, to further divide the remaining work. The splitter is called on a running task by an idle thread during a steal operation. The task and its splitter are concurrent and must be carefully managed as they both need to access shared data structure. The programmer is held responsible for writing correct task and splitter codes. To help him, the X-KAAPI runtime ensures that only one thief performs splitter concurrently with the task execution. It allows for simple and efficient synchronization protocols. Moreover, for applications developers, a set of higher parallel algorithms, like those of the STL [21], are proposed on top of the adaptive task model. Next section focuses on the parallel *foreach* algorithm.

2.5 Adaptive tasks for parallel loops

Following the OpenMP `parallel for` directive, X-KAAPI proposes a parallel loop function called `kaapic_foreach`, which is used in the backend of our X-KAAPI compiler [15].

A call to `kaapic_foreach` creates an adaptive task that iterates through the input interval $[first, last)$ to apply a functor (the loop body). The initial interval is partitioned in p slices, one slice reserved to each available core. When a thread calls the splitter to obtain work from the adaptive task, it grabs the reserved slice if available. The splitter returns an adaptive task that calls the functor for each iteration of the slice.

If the initial slice is not longer available, the splitter tries to split the interval $[b_t, e)$ corresponding to the iteration that remains to be process at time t. Thanks to the concurrency level guaranteed by the scheduler, a Dijkstra's like protocol ensures coherent split of interval while task iterates. The aggregation protocol is able to process k steal requests at once. The main thief tries to split $[b_t, e)$ into $k+1$ equal slices, leaving one slice for the victim. Then, for each of the k requests, the thief returns approximately the same amount of work for balancing purpose [20].

3 Benchmarks

This section presents a synthetic selection of three benchmarks to compare X-KAAPI performance with respect to three parallel programming models: fork-join model, parallel loops and data flow tasks.

The multicore platform used in this section is a 48 cores AMD Magny Cours platform with 256GBytes of main memory. Each core frequency is 2.2Ghz. The machine has 8 NUMA nodes. Each node has 6 cores sharing a L3 cache of 5 MBytes. Reported times are averaged over 30 runs.

3.1 Task creation time

This section compares the overhead of task creation and execution with respect to the sequential computation. The experiment evaluates the time to execute the X-KAAPI program of figure 1 for computing the 35-th Fibonacci number. The program recursively creates tasks without any data flow dependency. X-KAAPI is compared with Intel Cilk+ (icc 12.1.2), Intel TBB-4.0 and OpenMP-3.0 (gcc-4.6.2). Fibonacci is a standard benchmark used by Cilk [8] and Intel TBB (part of TBB-4.0 source code). Sequential time is $0.091s$. Figure 1 reports times using 1, 8, 16, 32 and 48 cores.

```
void fibonacci(long* result,
               const long n)
{
  if (n<2)
    *result = n;
  else
  {
    long r1,r2;
#pragma kaapi task write(&r1)
    fibonacci( &r1, n-1 );
    fibonacci( &r2, n-2 );
#pragma kaapi sync
    *result = r1 + r2;
  }
}
```

a. X-KAAPI Benchmark using KaCC

#cores	Cilk+	TBB	Kaapi	OpenMP
1 (slowdown:1)	1.063 (x 11.7)	2.356 (x 26)	0.728 (x 8)	2.429 (x27)
8	0.127	0.293	0.094	*51.06*
16	0.065	0.146	0.047	*104.14*
32	0.035	0.072	0.024	*(no time)*
48	0.028	0.049	0.017	*(no time)*

b. Time (second)

Figure 1: Fibonacci micro benchmark. Sequential time is 0.091s for Fibonacci 35.

Benchmark sources for OpenMP or TBB are not listed but they create exactly the same number of tasks and synchronization points. TBB has more overhead with respect to the sequential computation (slowdown of about 26) in comparison to X-KAAPI (slowdown of 8). This overhead can easily be amortized by increasing the task granularity, but at the expense of increasing the critical path, thus reducing the available parallelism [8]. OpenMP (gcc 4.6.2) performs poorly: the grain is too fine and OpenMP cannot speed up the computation. Computation was stopped on 32 and 48 cores after 5 minutes. The relatively good time of OpenMP with 1 core is due to an artifact of the libGOMP runtime: for one core, task creation is degenerated to a standard function call.

3.2 Data flow Cholesky factorization

The Cholesky factorization is an important algorithm in dense linear algebra. This section reports performances of the block version `PLASMA_dpotrf_Tile`

of PLASMA 2.4.2 [4]. On a multicore architecture, PLASMA relies on the runtime QUARK [22] to manage data flow tasks. QUARK only supports a subset of the functionalities offered by X-KAAPI. Thus, we have ported QUARK on top of X-KAAPI to produce a binary compatible QUARK library, which is linked with PLASMA algorithms for X-KAAPI experiments. Figure 2 reports the performances (GFlop/s) with respect to the matrix size

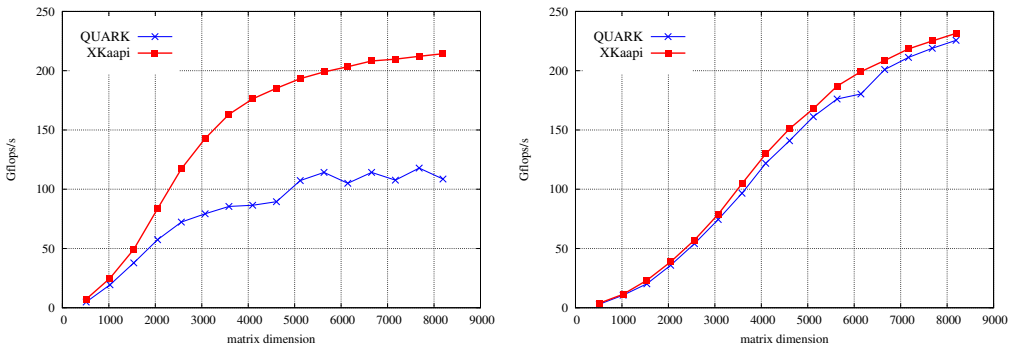

Figure 2: Gflops on Cholesky algorithm with QUARK and X-KAAPI. Tile size of $NB = 128$ (left) or $NB = 256$ (right).

on 48 cores. QUARK implements a centralized list of ready tasks, with some heuristics to avoid accesses to the global list. For fine grain tasks ($NB = 128$) and due to a contention point to access the global list, X-KAAPI outperforms QUARK. We can expect this contention point to become more severe as the core number increases with next generation machines, affecting PLASMA performance. When the grain increases, X-KAAPI remains better but the difference decreases because of the relatively small impact of task management with respect to the whole computation. One can also note that increasing the grain size reduces the average parallelism and limits the speedup. For matrix of size 3000, the performance for $NB = 128$ reaches almost $150GFlops$, while for $NB = 256$, it drops to about $75GFlops$.

3.3 Parallel independent loops

We compare OpenMP/GCC 4.6.2 parallel loop using static and dynamic schedule against X-KAAPI (`kaapic_foreach` version). Figure 3 reports speedups ($T_{seq} = 386s$) of the two parallel loops of the EUROPLEXUS application (next section). Both OpenMP static and dynamic schedule have the same performances. Globally, OpenMP and X-KAAPI speedups are very close, but X-KAAPI outperforms OpenMP past 25 cores. The same cores were used by X-KAAPI and OpenMP by binding threads to cores using an affinity mask.

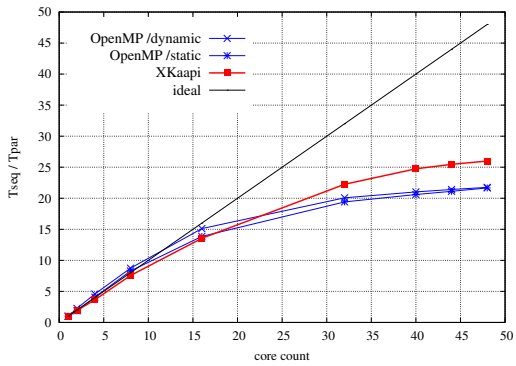

Figure 3: Comparison of parallel loop speedup

4 EUROPLEXUS

EUROPLEXUS is a computer program for the simulation of fluid-structure systems under transient dynamic loading. The code is co-owned since 1999 by French CEA and European Commission (Joint Research Center, Institute for the Protection ans Security of the Citizen) and jointly developed through a consortium also involving EDF (French national electricity board) and ONERA (French aerospace research labs). EUROPLEXUS uses finite elements, SPH particles or discrete elements to model structures and finite elements, finite volumes or SPH particles to model fluids. EUROPLEXUS is dedicated to simulating the consequences of extreme loadings such as explosion and impacts, with strong coupling between structures and fluids. Time integration is explicit (central difference schemes for structures and explicit Euler for fluids) and about 140 geometric elements are available, along with about 100 material models and about 50 kinds of kinematic connexions between entities, such as unilateral contact, fluid structure links for both conformant meshes and immersed boundaries or various kinds of constrained motions. To avoid non-physical parameters throughout the kinematic constraints enforcement procedure, Lagrange multipliers are used to compute link forces, yielding the need for linear system solvers alongside the classical explicit solution process.

The source code is thus complex (600.000 lines of Fortran) and many algorithms are involved simultaneously within classical simulations. However, two main kinds of algorithmic tasks accounts for 70% of a common EUROPLEXUS execution: 1/ independent parallel loops for nodal force vector evaluations and kinematic link detection; 2/ sparse Cholesky factorization of the so-called **H** matrix, obtained from the condensation of dynamic equilibrium equations onto Lagrange multipliers, in a Skyline representation (the cost of following triangular system solutions being neglected).

Practically, three main algorithms are considered for subsequent examples: representative of classical EUROPLEXUS simulations in structural domain. They are named from the Fortran procedure used in EUROPLEXUS to perform the task:

1. LOOPELM: independant loop on finite elements to compute nodal internal forces from local mechanical behaviour,
2. REPERA: independant loop to sort candidates for *node_to_facet* unilateral contact,
3. CHOLESKY: perform Cholesky factorization of a symetric positive semi-definite matrix.

The distribution of execution times between these three algorithms varies with the simulation step and with the considered instance. In this paper we focus on two simulation scenarios. The first one, called MEPPEN, consists in the crash of a large steel missile on a perfectly rigid wall. The second one, called MAXPLANE, consists in the impact of a ice projectile on a composite plate. Due to physics and modelling, these two instances provide very different repartitions of time among the considered algorithms:

- MEPPEN is characterized by large structural strains, strongly non-linear behaviour and multiple contacts as the missile undergoes dynamic buckling: time is then mainly split between LOOPELM, with large ratios between finite elements, and REPERA,
- MAXPLANE is characterized by a modelling of the composite plate plies using 3D finite elements, with contact conditions between each plies, so that the size and filling of the **H** matrix are close to those of the system stiffness matrix: the solution procedure is then strongly dominated by the condensed system solution, and then by the CHOLESKY algorithm.

4.1 LOOPELM and REPERA loops

Figure 4 details the X-KAAPI speedups for the MEPPEN (left) and MAXPLANE (right) instances. On the smallest instance, MEPPEN, the LOOPELM has limited speedup due to its memory intensive character. REPERA is more computation intensive leading to a good speedup.

4.2 Sparse Cholesky factorization

The sparse Cholesky factorization (LDL^t) represents about 60% of the execution time for the MAXPLANE instance. The numerical scheme requires to factor and solve a linear system at each time step. The linear system is sparse and its size and density depend on the interactions in the simulation.

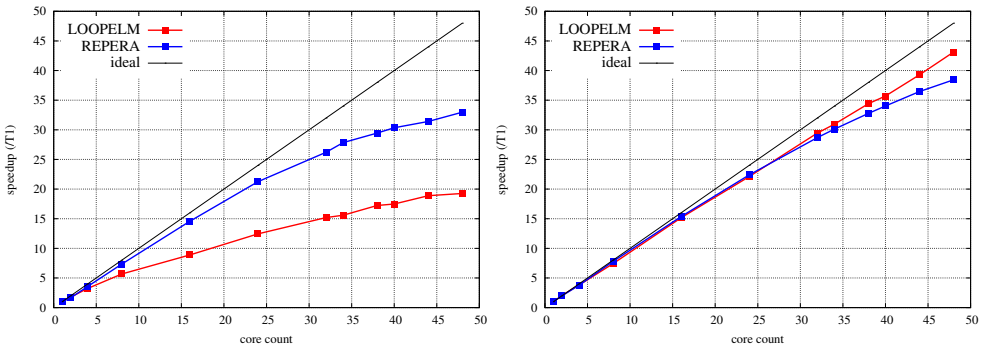

Figure 4: Speedups of LOOPELM and REPERA on MEPPEN and MAXPLANE.

```
1  for (k = 0; k < N; k += BS)
2  {
3    potrf(k, &sli);
4    for (m = k + BS; m < N; m += BS)
5    {
6      if (is_empty(m, k, &sli)) continue;
7      trsm(k, m, &sli);
8    }
9    for (m = k + BS; m < N; m += BS)
10   {
11     if (is_empty(m, k, &sli)) continue;
12     syrk(k, m, &sli);
13     for (n = k + BS; n < m; n += BS)
14     {
15       if (is_empty(n, k, &sli)) continue;
16       if (is_empty(m, n, &sli)) continue;
17       gemm(k, m, n, &sli);
18     }
19   }
20 }
```

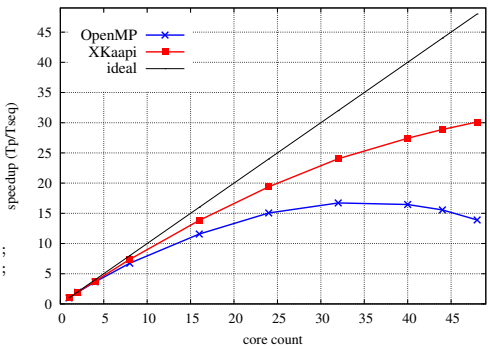

Figure 5: Sequential sparse Cholesky code. Speedups of X-KAAPI vs OpenMP.

The pseudo sequential code is sketch in figure 5. Variable `sli` is the skyline representation of the sparse matrix to factorize. The function calls `potrf`, `trsm`, `syrk` and `dgemm` at lines 3, 7, 12, and 17 are pseudo blas functions with `sli` the skyline matrix parameter and k, n, m the indexes delimiting the block to process. All these calls create tasks in the X-KAAPI version. Only calls at line 7, 12 and 17 create tasks in OpenMP.

In the X-KAAPI version, these indexes serves as defining memory accesses to compute dependencies. OpenMP parallelization implies synchronization between tasks in order to satisfy data flow dependencies. So, `#pragma omp taskwait` directives have to put after lines 8 and 19. As noted by [13], the parallel data flow version only specifies tasks with access modes, without explicit synchronizations.

Figure 5 reports speedup using a matrix that appear during the MAXPLANE simulation. The dimension of the matrix is 59462 with 3.59% of non zero elements. We looked for the best block size for this experiment:

$BS = 88$. The sequential time is $47.79s$. X-KAAPI version outperforms OpenMP (gcc-4.6.2) version, for the same reasons as for the dense Cholesky factorization [13].

4.3 Overall gains of EUROLEXUS

Figure 6 reports the performances of the parallel version with respect to the sequential code. The left bar in the histograms represents sequential time decomposition with respect to the time of each algorithm presented above. The Amdhal's law applies: we have started parallelization of the remaining sequential part.

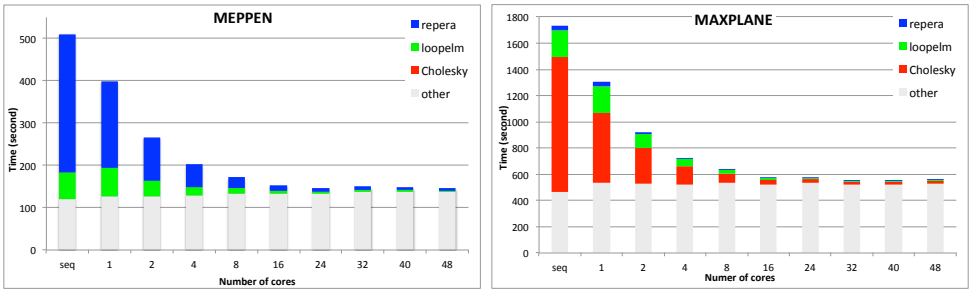

Figure 6: Overall gains of EUROPLEXUS with X-KAAPI.

5 Related work

Kaapi [10] was designed in our group after the preliminary work on Athapascan [9, 5]. X-KAAPI keeps definition of access mode to compute data flow dependencies between a sequence of tasks. StarSs/SMPSs [2], QUARK [22], StarPU [1] follow the same design. Differences are in the kind of access mode and the memory region shape that is defined : StarSs/SMPSs, QUARK have similar access mode and consider unidimensional array. QUARK has an original `scratch` access mode to reuse thread specific temporary data. StarPU [1] has a more complex way to split data and define sub-view of a data structure. X-KAAPI has direct support for multi-dimensional arrays.

The data flow task model is flat in StarSs/SMPSs, QUARK and StarPU while X-KAAPI allows recursive task creation. The fork-join parallel paradigm is only supported by X-KAAPI, Intel TBB [19], Cilk [8] and Cilk+ (Intel version of Cilk). The X-KAAPI performance for fine grain recursive applications is equivalent, or even better, than Cilk+ and Intel TBB that only allow independent task creations. In TBB, Cilk or X-KAAPI task creation is several order of magnitude less costly than in StartSs/SMPSs, QUARK or StarPU. QUARK and StarPU cannot scale well due to their central list scheduling. SMPSs seems to support a more distributed scheduling.

X-Kaapi has a unique model of adaptive task that allow a runtime adaptation of task creation when resources are idle. The OpenMP libGOMP runtime implements a threshold heuristic that limits task creation when the number of tasks is greater than 64 times the number of threads. It can limit the parallelism of the application and thus performance cannot be guaranteed like with a workstealing algorithm. TBB, with autopartitionner heuristic, is able to limit the number of tasks without, a priori, limit the parallelism of the application.

Intel TBB, Cilk+, OpenMP and X-Kaapi support parallel loop which are not present in StarSs/SMPSs, QUARK or StarPU. Our comparison with OpenMP/GCC 4.6.2 shows that for benchmarked instances and applications, scheduling strategy is not an important feature.

6 Conclusions and future directions

This paper introduced the X-Kaapi multi paradigm parallel programming model. Experiments highlighted that for each paradigm specific benchmark, X-Kaapi reaches a similar or better performance than the reference software for this paradigm. We also compare OpenMP and X-Kaapi on the industrial code EUROPLEXUS. If for the parallel loop parallelism, X-Kaapi and OpenMP show an equivalent performance (with better scalability for X-Kaapi), for data flow tasks the OpenMP parallel model imposes synchronizations that limits the speedup. This overhead experienced with our sparse Cholesky factorization, was already spotted in [13] on dense linear algebra factorizations.

This X-Kaapi evaluation draws two interesting conclusions: 1/ the OpenMP dynamic and static schedulers, which comes from historical design choices, would benefit from being unified. Intel TBB only proposes a dynamic scheduler; 2/ a (macro) data flow task model supporting recursivity can be efficiently implemented and be competitive with a simple fork-join model.

Ongoing work focuses on our compiler infrastructure, and to integrate our multi-CPUs multi-GPUs support [12] and distributed memory architecture support [10].

Acknowledgment

This work has been supported by CEA and by the ANR Project 09-COSI-011-05 Repdyn.

References

[1] Agullo, E., Augonnet, C., Dongarra, J., Ltaief, H., Namyst, R., Roman, J., Thibault, S., Tomov, S.: Dynamically scheduled Cholesky fac-

torization on multicore architectures with GPU accelerators. In: Symposium on Application Accelerators in High Performance Computing (SAAHPC). Knoxville, USA (2010)

[2] Badia, R.M., Herrero, J.R., Labarta, J., Pérez, J.M., Quintana-Ortí, E.S., Quintana-Ortí, G.: Parallelizing dense and banded linear algebra libraries using smpss. Concurr. Comput. : Pract. Exper. 21, 2438–2456 (2009)

[3] Blumofe, R.D., Leiserson, C.E.: Space-efficient scheduling of multithreaded computations. SIAM J. Comput. 27, 202–229 (1998)

[4] Buttari, A., Langou, J., Kurzak, J., Dongarra, J.: A class of parallel tiled linear algebra algorithms for multicore architectures. Parallel Comput. 35, 38–53 (2009)

[5] Dumitrescu, B., Doreille, M., Roch, J.L., Trystram, D.: Two-dimensional block partitionings for the parallel sparse cholesky factorization. Numerical Algorithms 16, 17–38 (1997)

[6] Duran, A., Ferrer, R., Ayguadé, E., Badia, R.M., Labarta, J.: A proposal to extend the openmp tasking model with dependent tasks. Int. J. Parallel Program. 37, 292–305 (June 2009)

[7] Fich, F.E.: New bounds for parallel prefix circuits. In: Proceedings of the fifteenth annual ACM symposium on Theory of computing. pp. 100–109. STOC '83, ACM, New York, NY, USA (1983)

[8] Frigo, M., Leiserson, C.E., Randall, K.H.: The implementation of the cilk-5 multithreaded language. SIGPLAN Not. 33, 212–223 (1998)

[9] Galilée, F., Roch, J.L., Cavalheiro, G.G.H., Doreille, M.: Athapascan-1: On-line building data flow graph in a parallel language. In: Proceedings of PACT'98. pp. 88–. PACT '98, IEEE Computer Society, Washington, DC, USA (1998)

[10] Gautier, T., Besseron, X., Pigeon, L.: KAAPI: A thread scheduling runtime system for data flow computations on cluster of multi-processors. In: Proceedings of PASCO'07. ACM, New York, NY, USA (2007)

[11] Hendler, D., Incze, I., Shavit, N., Tzafrir, M.: Flat combining and the synchronization-parallelism tradeoff. In: Proceedings of the 22nd ACM SPAA. ACM, New York, NY, USA (2010)

[12] Hermann, E., Raffin, B., Faure, F., Gautier, T., Allard, J.: Multi-GPU and Multi-CPU Parallelization for Interactive Physics Simulations. In: EUROPAR 2010. Ischia Naples, Italy (2010)

[13] Kurzak, J., Ltaief, H., Dongarra, J., Badia, R.M.: Scheduling dense linear algebra operations on multicore processors. Concurr. Comput. : Pract. Exper. 22, 15–44 (2010)

[14] Le Mentec, F., Danjean, V., Gautier, T.: X-Kaapi C programming interface. Tech. Rep. RT-0417, INRIA (2011)

[15] Le Mentec, F., Gautier, T., Danjean, V.: The X-Kaapi's Application Programming Interface. Part I: Data Flow Programming. Tech. Rep. RT-0418, INRIA (2011)

[16] Lee, E.A.: The problem with threads. Computer 39, 33–42 (2006)

[17] OpenMP Architecture Review Board: http://www.openmp.org (1997-2008)

[18] Quinlan, D.J., *et al.*: Rose compiler project, `http://www.rosecompiler.org`

[19] Reinders, J.: Intel threading building blocks. O'Reilly & Associates, Inc., Sebastopol, CA, USA, first edn. (2007)

[20] Tchiboukdjian, M., Gast, N., Trystram, D., Roch, J.L., Bernard, J.: A Tighter Analysis of Work Stealing. In: The 21st International Symposium on Algorithms and Computation (ISAAC). No. 6507, Springer, Jeju Island, Korea (2010)

[21] Traore, D., Roch, J.L., Maillard, N., Gautier, T., Bernard, J.: Deque-free work-optimal parallel STL algorithms. In: EUROPAR 2008. Springer-Verlag, Las Palmas, Spain (2008)

[22] YarKhan, A., Kurzak, J., Dongarra, J.: Quark users' guide: Queueing and runtime for kernels. Tech. Rep. ICL-UT-11-02, University of Tennessee (2011)

RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée

655 avenue de l'Europe Montbonnot

38334 Saint Ismier Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399