

LU, QR and Cholesky factorizations: Programming Model, Performance Analysis and Optimization Techniques for the Intel Knights Landing Xeon Phi

DUFOYER Benjamin, CHEVALIER Arthur

1 INTRODUCTION

Today's modern computers have a wide variety of heterogeneous compute resources, ranging from multicore CPUs to GPUs and to face up to the increase quantity of numerical informations and modern systems complexity, it is necessary to enhance the structures that manage this task. The main thing of the heterogeneous system is the capacity of every resources to work in the same time. A good task scheduling is paramount to increase performance on this type of system. The biggest problem is to chop every task to execute in the same time while avoiding the competitor accesses. With the multiplication of core number, the developers must change their way of programming in order to slice the problem to be able to parallelize their application with high-performance. Intel released a new Xeon PHI called KNL (Knight Landing) which has up to 72 cores and this paper is about testing a library called MAGMA in this new highly parallelizable platform. In the matrix factorizations algorithm the MAGMA library provide a way to slice the matrix to ensure that all the tasks will run at the same time and with minimal idle time. The library also use a way to divide all the cores of the KNL in order to simulate multiple devices to use. The article will compare performances over three different matrix factorization algorithms: LU, QR and Cholesky.

Presentation of KNL, KNL is the new version of Xeon Phi. He is on the Intel Family of Many Integrated Core architecture (MIC). He is available in two form, coprocessor and host processor. He will be built using up to 72 cores with four threads per core and use LGA 3647 socket. Each core are engrave at 14nm. It supports for up 384Gb of DDR4 RAM and for up 16Gb of MCDRAM. The MCDRAM is a innovation for the Xeon, it's a 3D-stacked DRAM. The favours of the RAM is the fast access to data (better than classical RAM) on the test, they use the MCDRAM in flat mode, so the MCDRAM have her own adress space and can be adress explicitly. The next innovation of the KNL is the addition of two 512-bit vectorial unit on each core. The KNL achieve a double precision theoretical peak of 2662 Gflops/s. The version used in the test is the selfhosted preproduction

Intel Xeon Phi card with 16Gb of MCDRAM, each core run at 1,3 GHz. It consists of 64 cores with 4 hyperthreads each. The KNL have different mod of utilisation. You can use in InterConnect mode, in this mode, all core are connected with his neighbor's. Other mode are Cluster Mode with a "All to All", quadrant and Sub-Numa. In the test, they use the Quadrant Mode to increase performance. The principe is simple, the card is cut into four part, every part have a unique Master and the core remaining are a slave.

2 CONTRIBUTION

The MAGMA (Matrix Algebra on GPU and Multi-core Architectures) project aims to develop a dense linear algebra library similar to LAPACK but for heterogeneous/hybrid architectures, starting with current "Multi-core and GPU" systems. The goal of MAGMA is to combine the strengths of different algorithms within a single framework. They want to design linear algebra algorithms and frameworks for hybrid manycore and GPUs systemes that can enable applications to fully exploit the power that each of the hybrid components offer. In practice MAGMA want to reduce the communication into every function to have a better parallelism.

3 ELEMENTS D'IMPLEMENTATION

MAGMA actually uses the BLAS libraries to compute the LU, QR and Cholesky matrix factorizations. The algorithm works as follow:

First we factorize the first row and column of the matrix and when we finish this task we update the trailing matrix on the GPU. To use parallelism on this algorithm the idea is to start to factorize the first line and column of the trailing matrix as soon as possible so when the GPU updated them. The factorization works on the CPU to maximize the performances. The BLAS2 library is used on the factorization of single vectors (column and row) and the BLAS3 is used on the update of the trailing matrix. We can also vary the depth level of the algorithm because we can have at the same time multiple updating task on the trailing matrix.

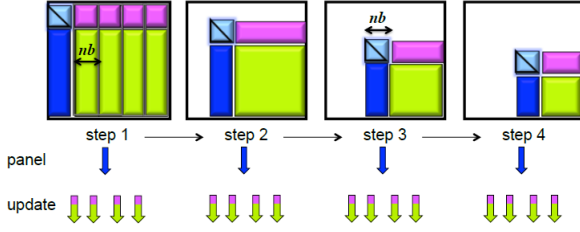


Figure 1. Two-phase implementation of a one-sided factorization

When the second factorization ended we launch a second update on the trailing matrix while the first is still running. The number of updating tasks is the depth level.

4 VALIDATION/EXPERIMENTATION

Three different configurations has been used for the performances tests:

- The first one configuration has two 18-core Intel Xeon E5-2697 (Broadwell) processors, running at 2,6 GHz. Each socket as 35 MiB of shared L3 cache, and each core has a private 3,5 MiB L2 and 448 KB L1 cache. The system is equipped with 52 GB of memory and the theoretical peak in double precision is 20,8 Gflop/s per core.
- The second system is equipped with two 8-core Intel Xeon E5-2670 (Sandy Bridge) processors and two Intel Xeon Phi 7120 (KNC) cards with 15,8 GB per card. She is equip with 61 core running at 1,23GHz. And achieving a double precision theoretical peak of 1180 Gflops/s. The KNC is the old version of Xeon Phi and don't have MCDRAM and vector unit.
- The last is a selfhosted preproduction Intel Xeon Phi Knights Landing (KNL) card, with 16 GB of MCDRAM used in flat mode, running at 1.3GHz and achieving a double precision at a theoretical peak of 2662 Gflop/s. It has 64 cores with 4 hyper-threads each.

The tests are performed over the three different factorizations algorithms with varying matrix size (from 2k to 40k cells for border size). For the LU factorization we can see that behind the 6k size the two KNL and the first one are far over the others with approximately twice the performance. The KNL reached its limit of performance at 28k with a peak of 1500Gflop/s and come behind the second configuration that reach the 1700Gflop/s with the 40k matrix. We can also see that the KNL cannot perform tests over 30k matrix because of its memory.

For the QR factorization we have the same schema, the third configuration get to follow the KNL until 8k matrix size but get overtaken with taller matrix. With medium matrix size the KNL is still leading with the third configuration like the LU factorization.

Finally for the Cholesky factorization the second config is ahead on small matrix (6k size) but the KNL leads the way until 16k matrix size and the last configuration (two KNC and one processor) get better performance on large matrix.

To sum up this different tests, the new Xeon Phi proposed by Intel obtain better performances than the older

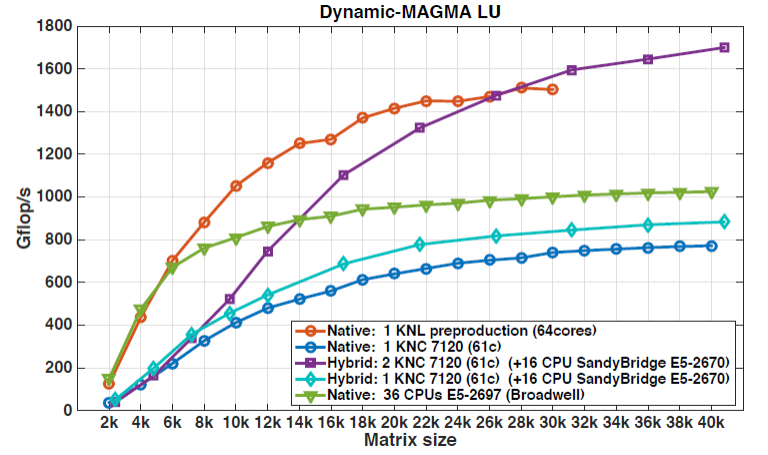


Figure 2. LU performances with MAGMA

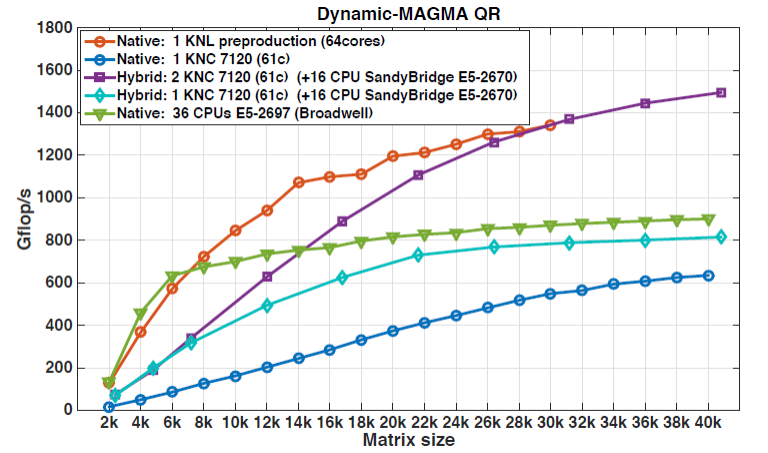


Figure 3. QR performances with MAGMA

version (KNC) but is not unbeatable on performance contest. We can see that it obtains remarkable performances on small matrix but we easily reached the limit of performance on medium matrix.

5 CONCLUSION

5.1 Discussion

5.2 Perspectives à court terme

At this time, the KNL is in testing phase. This paper is about testing matrices factorizations on it and a lot of others performances tests will appear soon.

5.3 Perspectives à long terme

In the long term, the KNL will be used in clusters for high performance processing

REFERENCES

[1]

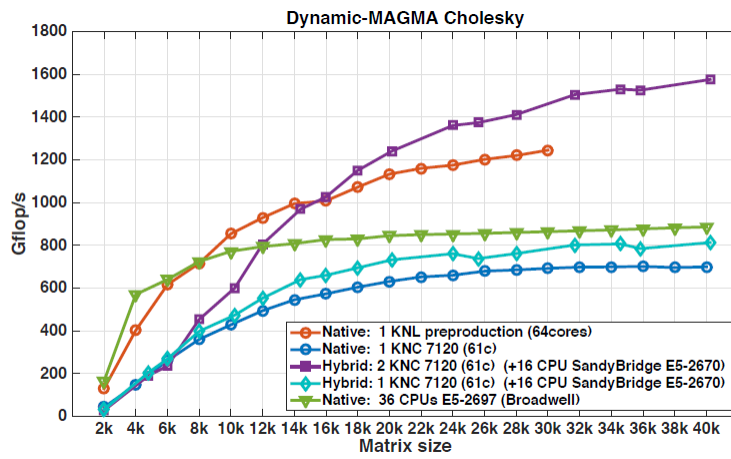


Figure 4. Cholesky performances with MAGMA