Performance predictions of multilevel communication optimal LU and QR factorizations on hierarchical platforms

Grigori, Laura and Jacquelin, Mathias and Khabou, Amal

2013

MIMS EPrint: **2013.11**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: http://eprints.maths.manchester.ac.uk/
And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

Performance predictions of multilevel communication optimal LU and QR factorizations on hierarchical platforms

Laura Grigori[1], Mathias Jacquelin[2], and Amal Khabou[3]

[1] INRIA Paris - Rocquencourt, Paris, France
laura.grigori@inria.fr
[2] Lawrence Berkeley National Laboratory, Berkeley, USA
mjacquelin@lbl.gov
[3] The University of Manchester, Manchester, UK
amal.khabou@manchester.ac.uk

Abstract. In this paper we study the performance of two classical dense linear algebra algorithms, the LU and the QR factorizations, on multilevel hierarchical platforms. We note that we focus on multilevel QR factorization, and give a brief description of the multilevel LU factorization. We first introduce a performance model called Hierarchical Cluster Platform (Hcp), encapsulating the characteristics of such platforms. The focus is set on reducing the communication requirements of studied algorithms at each level of the hierarchy. Lower bounds on communication are therefore extended with respect to the Hcp model. We then present a multilevel QR factorization algorithm tailored for those platforms, and provide a detailed performance analysis. We also provide a set of performance predictions showing the need for such hierarchical algorithms on large platforms.

Keywords: QR, LU, exascale, hierarchical platforms.

1 Introduction

Numerical algorithms and solvers play a crucial role in scientific computing. They lie at the heart of many applications and are often key to performance and scalability. Due to the ubiquity of multicore processors, solvers should be adapted to better exploit the hierarchical structure of modern architectures, where the tendency is towards multiple levels of parallelism. Thus with the increasing complexity of nodes, it is important to exploit these multiple levels of parallelism even within a single compute node. For this reason, classical algorithms need to be revisited so as to fit modern architectures that expose parallelism at different levels in the hierarchy. We believe that such an approach is mandatory in order to exploit upcoming hierarchical exascale computers at their full potential.

Studying the communication complexity of linear algebra operations and designing algorithms that are able to minimize communication is a topic that has received an important attention in the recent years. The most advanced approach in this context assumes one level of parallelism and takes into account the computation, the volume of communication, and the number of messages exchanged

along the critical path of a parallel program. In this framework, the main previous theoretical result on communication complexity is a result derived by Hong and Kung in the 80's providing lower bounds on the volume of communication of dense matrix multiplication for sequential machines [1]. This result has been extended to parallel machines [2], to dense LU and QR factorizations (under certain assumptions) [3], and then to basically all direct methods in linear algebra [4]. Given an algorithm that performs a certain number of floating point operations, and considering the memory size, the lower bounds on communication are obtained by using the Loomis-Whitney inequality, as for example in [2, 4]. While theoretically important, these lower bounds are derived with respect to a performance model that supposes a memory hierarchy in the sequential case, and P processors without memory hierarchy in the parallel case. Such a model is not sufficient to encapsulate the features of modern hierarchical architectures.

On the practical side, several algorithms have been introduced recently [5, 6, 7, 8]. Most of them propose to use different reduction trees depending on the hierarchy. However, the focus is set on reducing the running time without explicitly taking communication into consideration. In [8], Dongarra et al. propose a generic algorithm implementing several optimizations regarding pipelining of computation, and allowing to select different elimination trees on platforms with two levels of parallelism. They provide insights on choosing the appropriate tree, a binary tree being for instance more suitable for a cluster with many cores, while a flat tree allows more locality and CPU efficiency. However, neither theoretical bounds nor cost analysis are provided in these studies. Moreover, even if cache-oblivious algorithms are natural good candidates for reducing communication requirements at every level, they are not good candidates for large parallel implementations. We thus focus on cache- and parallelism- aware algorithms.

In the first part of this paper we introduce a performance model that we refer to as the Hierarchical Cluster Platform (Hcp) model. Provided that two supercomputers might have different communication topologies and different compute nodes with different memory hierarchies, a detailed performance model tailored for one particular supercomputer is likely to not reflect the architecture of another supercomputer. Hence the goal of our performance model is to capture the main characteristics that influence the communication cost of peta- and exa- scale supercomputers which are based on multiple levels of parallelism and memory hierarchy. We use the proposed Hcp model to extend the existing lower bounds on communication for direct linear algebra, to account for the hierarchical nature of present-day computers. We determine the minimum amount of communication that is necessary at every level in the hierarchy, in terms of both number of messages and volume of communication. Moreover, to the best of our knowledge, there is currently no algorithm targeting hierarchical platforms with more than two levels, nor any lower bound on communication for such platforms.

In the second part of the paper we introduce a multilevel algorithm for computing the QR factorization (*ML-CAQR*) that is able to minimize the communication at each level of the hierarchy, while performing a reasonable amount of extra computations. We note that we have also developed two multilevel algo-

rithms for the LU factorization (1D-*ML-CALU* and 2D-*ML-CALU*). However we restrict our study to the QR factorization here. We refer interested readers to the technical report [9] for more details about the multilevel LU algorithms. These recursive algorithms rely on their corresponding 1-level algorithms (resp. *CAQR* and *CALU*) as their base case. Indeed, *CAQR* and *CALU* are known to attain the communication lower bounds in terms of both bandwidth and latency with respect to the simpler one level performance model.

2 Background: the QR factorization

The QR factorization of an m-by-n matrix is a widely used algorithm, be it for orthogonalizing a set of vectors or for solving least squares problems with m equations and n unknowns, where $m \geq n$. It is known to be an expensive $mn^2 + 1/3n^3 + O(n^2)$, but very stable factorization. It is thus crucial to optimize its performance. The algorithm decomposes an m-by-n matrix A into two matrices Q and R such that $A = QR$, where Q is an m-by-m orthogonal matrix, while the m-by-n matrix R is upper triangular.

The QR factorization is obtained by applying a sequence of m-by-m unitary orthogonal transformations on the input matrix A. An unitary transformation U_i introduces some zeros below the diagonal in the current updated matrix. The two basic transformations are Givens rotations and Householder reflections. A Givens rotation introduces a single zero while a Householder reflection zeroes out every element below the diagonal. Using Givens rotations, disjoint pairs of rows can be processed concurrently. Householder reflections, though not displaying the same parallelism, are less computationally expensive.

Tree-based algorithms intent to benefit from both methods. Householder transformations are applied on local domains, or tiles, before getting eliminated two-by-two in a Givens-like approach. Communication Avoiding QR (*CAQR*) [3] belongs to this category, and organizes the computations so as to match the lower bounds on communication introduced in [4]. After $min(m, n)$ transformations, the resulting R factor is stored in place in the upper triangular part of matrix A while the matrix Q is assumed to be implicitly stored in the lower triangular part using the compact WY representation for Householder reflections [10]. If needed, Q can be retrieved at the cost of extra computations by computing $Q = I - YTY^T$.

3 Toward a realistic Hierarchical Cluster Platform model (Hcp)

The focus of this study is set on hierarchical platforms running HPC applications and displaying increasingly deeper hierarchies. Such platforms are composed of two kinds of hierarchies: (1) a network hierarchy composed of interconnected network nodes, stacked on top of a (2) compute nodes hierarchy [11]. This compute hierarchy can be composed for instance of shared memory NUMA multicore

nodes. Moreover, on most modern supercomputers, compute nodes are often grouped into *drawers* displaying higher local communication speeds. Such drawers typically belong to the network hierarchy, which is clearly not only a router hierarchy.

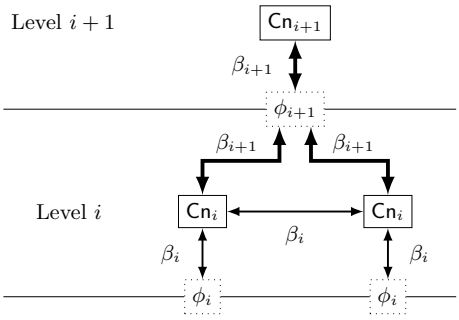


Fig. 1. Components of a level i in the HCP model.

The HCP model considers such platforms with l levels of parallelism, and uses the following assumptions. Level 1 is the deepest level in the hierarchy, where actual processing elements are located (for example cores). Each of these processing elements has its own local memory of size M_1 and a computing speed γ. A compute node of level $i+1$, denoted as Cn_{i+1} on Figure 1, is formed by P_i compute nodes of level i (two nodes in our example). The total number of processing elements of the entire platform is $P = \prod_{i=1}^{l} P_i$, while the total number of compute nodes of level i is $P_i^* = \prod_{j=i}^{l} P_j$. We let $M_i = M_1 \cdot \prod_{j=1}^{i-1} P_j$ be the aggregated memory size of a node of level $i > 1$.

The network latency α_i and the inverse bandwidth β_i apply throughout an entire level i. Moreover, we assume that generally, the higher in the hierarchy, the more expensive communication costs.

We also consider a message aggregation capacity ϕ_i at each level of the hierarchy, which determines the actual number of messages required to send a given amount of data. We refer to the number of messages sent at level i as S_i, and to the exchanged volume of data as W_i. $\bar{S}_i = S_i \cdot \alpha_i$ is the associated latency cost, while $\bar{W}_i = W_i \cdot \beta_i$ is the bandwidth cost. These notations will be used throughout the rest of the paper.

For the sake of simplicity in both algorithm description and cost analysis, we assume the P_i compute nodes of level i to be virtually organized along a 2D grid topology, that is $P_i = P_{r_i} \times P_{c_i}$ (note that any topology could be mapped onto a 2D grid).

We note that the model makes abstraction of the detailed architecture of a compute node or the interconnection topology at a given level of the hierarchy. Hence such an approach has its own limitations, since the predicted performance might not be accurate. However, while keeping the model tractable, this model better reflects the actual nature of supercomputers than the one level model assumed so far, and helps to understand the communication bottlenecks of common linear algebra operations. We also note that our model does not apply

to platforms with heterogeneity in the processing elements such as GPU and multi-GPU clusters.

Communicating under the Hcp model We now describe how communication happens in the Hcp model, and how messages flow through the hierarchy. We assume that if a compute node of level i communicates, all of its lower level nodes participate. Hence if some date has to be sent over the network, it first has to be collected from all the cores available on one node. We denote as ***counterparts*** of a compute node of level i all the nodes of level i lying in remote compute nodes of level $i+1$ and having the same local coordinates. We therefore have the relation $W_i = W_{i+1}/P_i$.

As an example, let us detail a communication of a volume of data W_i taking place between two nodes of level i. A total of P/P_i^* processing elements of level 1 are involved. Each has to send a chunk of data $W_1 = W_i P_i^*/P$. Since this amount of data has to fit in memory, we obviously have $\forall i, M_1 \geq W_1 = W_i P_i^*/P$. These blocks are transmitted to the level above in the hierarchy, i.e. to level 2. A compute node of level 2 has to send a volume of data $W_2 = P_1 W_1$. Since the aggregation capacity at level 2 is ϕ_2, this requires (W_2/ϕ_2) messages. The same holds for any level k such that $1 < k \leq i$, where data is forwarded by sending(W_k/ϕ_k) messages. We therefore have the following costs:

$$\bar{W}_k = \frac{W_i P_k^*}{P_i^*} \cdot \beta_k, \qquad \bar{S}_k = \frac{W_k}{\phi_k} \cdot \alpha_k = \frac{W_i P_k^*}{\phi_k P_i^*} \cdot \alpha_k.$$

This "regular" communication pattern is often encountered in HPC applications, the main target of the Hcp model, and is simpler than a purely heterogeneous pattern (which could be encountered in grid environments for instance). Moreover, this organization allows to aggregate data at the algorithm level rather than relying on the actual network topology.

It is interesting to note that the Hcp model allows to model several types of networks, depending on their aggregation capacity. We defined the three following network types to demonstrate Hcp versatility:

- *Fully-pipelined networks*, aggregating all incoming messages into a single message. This case is ensured whenever $\phi_i \geq P_{i-1} W_{i-1}$. Since M_i is the size of the largest message sent at level i, we assume $\phi_i = M_i$. We also assume that all levels below are themselves fully-pipelined. Therefore, the aggregation capacity becomes $\phi_i = M_i = P_{i-1}\phi_{i-1}$.
- *Aggregating networks*, aggregating data up to volume of $\phi_i < M_i$ before sending a message.
- *Forward networks*, where messages coming from lower levels are simply forwarded. For a given level i, it is required that $\phi_i = \phi_{i-1}$: when each sub-node from level $i-1$ sends S_{i-1} messages, the number of forwarded messages is $\bar{S}_i = P_{i-1}\bar{S}_{i-1}$.

Based on the two extreme cases, we assume the aggregation capacity ϕ_i to satisfy $\phi_{i-1} \leq \phi_i \leq P_{i-1}\phi_{i-1}$.

An example of hierarchical platform modeled by Hcp Consider a distributed memory platform composed of D drawers having N compute nodes apiece. Let each node be a NUMA shared memory machine, with P processors. Within a node, each socket is connected to a local memory bank of size M, thus leading to a total shared memory of size $M \times P$ per node.

Within a drawer, nodes are interconnected with high speed interconnect such as fiber optics, whereas drawers are connected with more classical copper links. Let inter-drawer communication bandwidth and latency respectively be W_{inter} and S_{inter}. Let intra-drawer communications have a bandwidth W_{d} and a latency S_{d}. For intra-node communications, we let W_{mem} (resp. S_{mem}) be the bandwidth (resp. latency) to exchange data with memory.

We model this platform in Hcp using three levels, with the following characteristics:

# Comp. nodes	Bandwidth	Latency	Memory	Agg. capacity
$P_1 = P$	$W_1 = W_{\text{mem}}$	$S_1 = S_{\text{mem}}$	$M_1 = M$	$\phi_1 = M$
$P_2 = N$	$W_2 = W_{\text{d}}$	$S_2 = S_{\text{d}}$	$M_2 = P_1 M_1$	$\phi_2 = M \cdot P_1$
$P_3 = D$	$W_3 = W_{\text{inter}}$	$S_3 = S_{\text{inter}}$	$M_3 = P_2 M_2$	$\phi_3 \leq M \cdot P_2$

The aggregation capacities are chosen as follows: (1) On such hierarchical platform, a processor is able to transfer, in one message, its entire local bank of memory to another processor within the same compute node. This is ensured by setting ϕ_1 to M. (2) A compute node can transfer its entire shared memory to a remote node in the same drawer in a single message. The aggregation capacity is therefore chosen as $\phi_2 = MP_1$. (3) Finally, at the topmost level, the interconnect generally does not allow for sending the global volume of data coming from all drawers using a single message. The aggregation capacity is thus chosen as $\phi_3 \leq MP_2$.

Hcp allows to model typical HPC platforms, giving communication details at each level of the hierarchy. The switch from a shared memory to a distributed memory environment is handled through the choice of the aggregation capacities.

Lower bounds on communication We now introduce lower bounds on communication at every level of the hierarchy. Lower bounds on communication have been generalized in [4] for direct methods of linear algebra algorithms which can be expressed as three nested loops. We refine these lower bounds under our hierarchical model. For matrix product-like problems, at least one copy of the input matrix has to be stored in memory: a compute node of level i thus needs a memory of $M_i = \Omega(n^2/P_i^*)$. Furthermore, the lower bound on latency depends on the aggregation capacity ϕ_i of the considered level i, where a volume $\bar{W}_i$ needs to be sent in messages of size ϕ_i. Hence the lower bounds on communication at level i:

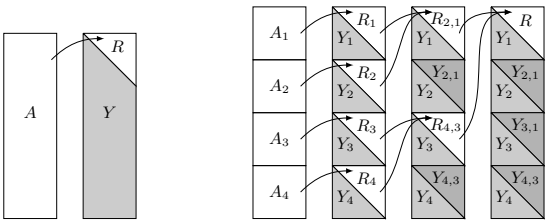
$$W_i \geq \Omega\left(\frac{\#flops}{\sqrt{memory}}\right) = \Omega\left(\frac{n^2}{\sqrt{P_i^*}}\right) \tag{1}$$

$$S_i \geq \Omega\left(\frac{W_i}{\phi_i}\right) = \Omega\left(\frac{n^2}{\phi_i\sqrt{P_i^*}}\right) \tag{2}$$

Note that, for simplicity, we expressed the bound on latency with respect to ϕ_i for each level i. Since we consider $\phi_1 = M_1$, the lower bound on latency for level 1 can also be expressed as $\bar{S}_1 = \Omega\left(\sqrt{P}\right)$.

4 Multilevel QR factorization

In this section we introduce *ML-CAQR*, a multilevel algorithm for computing the QR factorization of a dense matrix A. This multilevel algorithm heavily rely on its 1-level communication optimal algorithm *CAQR*, and can be seen as a recursive version of this algorithm. *ML-CAQR* recursive layout naturally allows for local elimination trees adapted to fit hierarchical platforms, thus reducing the communication needs at each level of the hierarchy. *ML-CAQR* is targeting large scale hierarchical platforms. The focus is set on keeping the communication requirements as low as possible at every level of the hierarchy, like *CAQR* on platforms with one level of parallelism.



(a) Single domain **(b)** Tree-based elimination

Fig. 2. Structure of the Householder reflectors

ML-CAQR, given in Algorithm 1, uses a recursive tree-based elimination scheme based on Householder reflections. As a tree-based algorithm, *ML-CAQR* stores the Householder reflectors in the lower triangular part of matrix A using a tree structure as in [3]. A small example is depicted on Figure 2, where a panel of matrix A is first split into four domains which are independently factored, then eliminated two by two. The resulting Householder reflectors should be applied following the same order to reflect the update of this panel.

At the topmost level of the hierarchy, *ML-CAQR* factors the entire input matrix A panel by panel. A panel is processed in multiple elimination steps following a tree-based approach. At the leaves of the tree, rectangular blocks are factored. The obtained R factors are then grouped two-by-two and eliminated in a sequence of elimination of size $2b_l$-by-b_l, where b_l is the panel size. Each factorization or elimination corresponds to a recursive call to *ML-CAQR* on the next lower level. After panel factorization, Householder reflectors are sent to remote compute nodes so as to update the trailing matrix using two recursive routines: *ML-Fact* and *ML-Elim*.

When called on two aggregated R factors, *ML-CAQR* and *ML-Elim* take this specific shape into account and do not perform any unnecessary computations.

Algorithm 1: *ML-CAQR*(A, m, n, r, P)

Input: Matrix A, m is the number of rows of A, n is the number of columns, r is the level of recursion, P is the current compute node
Output: Factored matrix with R in the upper triangular part and the Householder reflectors Y in the lower triangular part

```
if r = 1 then
    Call CAQR(A, m, n, b_1, P)
else
    for kk ← 1 to n, with step of b_r do
        for Compute nodes p ← 1 to P_{r_r} in parallel do
            h_p ← max(b_r, (m − kk + 1)/P_{r_r})
            if kk + (p − 1)h_p ≤ n then
                panel ← A(kk + (p − 1)h_p : kk + p · h_p − 1, kk : kk + b_r − 1)
                Call ML-CAQR(panel, h_p, b_r, r − 1, p)
        if There are multiple R factors then
            for j ← 1 to log P_{r_r} do
                Nodes (p_source, p_target) used to perform the elimination.
                Send local b_r-by-b_r to the remote node p_target
                Stack two b_r-by-b_r upper triangular matrices in RR
                Call ML-CAQR(RR, 2b_r, b_r, r − 1, p_source)
                Call ML-CAQR(RR, 2b_r, b_r, r − 1, p_target)
        for Compute nodes p ← 1 to P_{r_r} in parallel do
            Broadcast Householder vectors along processor row
            for Compute node rp ← 2 to P_{c_r} on same row as p do
                Call ML-Fact(r − 1, rp)
        if There are multiple R factors then
            for j ← 1 to log P_{r_r} do
                Nodes (p_source, p_target) used to perform the elimination.
                for Nodes rp ← 2 to P_{c_r} on same row as p_source in parallel do
                    Remote node rp_target is on same row as p_target and same column than rp
                    rp sends its local A to rp_target
                    Call ML-Elim(r − 1, rp)
                    Call ML-Elim(r − 1, rp_target)
```

However, for the sake of simplicity, this special case is not taken into account in Algorithm 1.

More formally, for each recursion level r, let b_r be the block size, and s be the internal computation step (it is incremented by b_r).

For each panel of size b_r, *ML-CAQR* proceeds as follows:

1. The panel is factored by using a reduction operation, where *ML-CAQR* is the reduction operator. With a binary tree, it processes as follows:
   (a) First, the panel is divided into P_{r_r} subdomains of size $(m - s + 1)/P_{r_r}$-by-b_r, which are recursively factored with *ML-CAQR* at level $r - 1$. At the deepest level, *CAQR* is called.
   (b) The resulting b_r-by-b_r R factors are eliminated two-by-two by *ML-CAQR* at level $r - 1$, requiring $\log P_{r_r}$ steps along the critical path.

The computation is redundantly performed on each pair of processors as it simplifies the communication pattern.

2. The current trailing matrix is then updated:
   (a) Householder reflectors in lower trapezoidal part of the panel have to be broadcasted along processor rows.
   (b) Updates corresponding to factorizations at the leaves of the tree are applied using the *ML-Fact* routine.
   ML-Fact broadcasts P_{r_r} blocks of Householder reflectors of size $(m-s+1)/P_{r_r}$-by-b_r from the column of nodes holding the current panel along rows of compute nodes. At the deepest level, the update corresponding to a leaf is applied as in *CAQR* (see [3]).
   (c) The updates due to the eliminations of the intermediate R factors are then applied to the trailing matrix using the *ML-Elim* procedure. Blocks of size b_r-by-$(n-s-b_r+1)/P_{c_r}$ are exchanged within a pair of compute nodes. At the lowest level, a partial update is locally computed before being independently applied onto each processing elements (similarly to *CAQR*).

5 Multilevel QR performance model

In this section, we provide cost analysis of *ML-CAQR* algorithm with respect to the HCP model. Two types of communication primitives are used, namely point-to-point and broadcast operations. To simplify the analysis, we define two recursive costs corresponding to these communication patterns.

In a *point-to-point communication*, a volume D is transferred between two compute nodes of level r. All compute nodes from level 1 to level $r-1$ below those two nodes of level r are involved, sending their local data to their respective counterparts in the remote node of level r. The associated communication costs are therefore:

$$\bar{W}_{\mathrm{P2P}}(1\dots r, D) = \sum_{k=1}^{r} \frac{D \cdot P_r^*}{P_k^*} \beta_k,$$

$$\bar{S}_{\mathrm{P2P}}(1\dots r, D) = \alpha_1 + \sum_{k=2}^{r} \frac{D \cdot P_r^*}{\phi_k P_k^*} \alpha_k.$$

A *broadcast operation* between P_{c_r} compute nodes of level r is very similar to point to point communication. However at every level, a participating node broadcasts its data to P_{c_r} counterparts. A broadcast can thus be seen as $\log P_{c_r}$ point-to-point communications.

We now review the global computation and communication costs of *ML-CAQR*. At each recursion level r, the current panel is factored by doing P_{r_r} parallel calls to *ML-CAQR*. Then, the resulting R factors are eliminated through $\log P_{r_r}$ successive factorizations of $2b_r$-by-b_r matrices formed by stacking up two upper triangular R factors. Once a panel is factored, the trailing matrix is updated. However, as the Householder reflectors are stored in a tree structure, the updates must be done in the same order as during panel factorizations. These

operations are recursively performed using *ML-Fact* for the leaves and *ML-Elim* for higher levels in the tree. The global recursive cost of *ML-CAQR* is composed of several contributions. We let:

- $T_{CAQR}(m,n,b,P)$ be the cost of factoring a matrix of size m-by-n with *CAQR* using P processors and a block size b.
- $T_{ML\text{-}CAQR}(m,n,b,P)$ be the cost of *ML-CAQR* on an m-by-n matrix using P processors and a block size b.
- $T_{\text{P2P}}(\text{levels}, \text{volume})$ be the cost of sending an upper triangular R factor within a panel of level r.
- $T_{ML\text{-}Fact}(m,n,b,P)$ be the cost of updating the trailing matrix to reflect factorizations at the leaves of the elimination trees.
- Finally, $T_{ML\text{-}Elim}(m,n,b,P)$ be the cost of applying updates corresponding to higher levels in the trees.

In terms of communication, *ML-Fact* consists in broadcasting Householder reflectors along process rows, while *ML-Elim* corresponds to $\log P_{r_r}$ point to point communications of trailing matrix blocks between pairs of nodes within a process column. Using these notations, the cost $T_{ML\text{-}CAQR}(m,n,b_r,P_r)$ of *ML-CAQR* can be expressed as,

$$\begin{cases} \sum_{s=1}^{n/b_r} \Big[T_{ML\text{-}CAQR}\left(\frac{m-(s-1)b_r}{P_{r_r}}, b_r, b_{r-1}, P_{r-1}\right) \\ \quad + \log P_{r_r} \cdot T_{\text{P2P}}(1\ldots r, \frac{b_r^2}{2}) \\ \quad + \log P_{r_r} \cdot T_{ML\text{-}CAQR}(2b_r, b_r, b_{r-1}, P_{r-1}) & \text{if } r > 1 \\ \quad + T_{ML\text{-}Fact}\left(\frac{m-(s-1)b_r}{P_{r_r}}, \frac{n-sb_r}{P_{c_r}}, b_{r-1}, P_{r-1}\right) \\ \quad + \log P_{r_r} \cdot T_{ML\text{-}Elim}\left(2b_r, \frac{n-sb_r}{P_{c_r}}, b_{r-1}, P_{r-1}\right) \Big] \\ \\ T_{CAQR}(m,n,b_1,P_1) & \text{if } r = 1 \end{cases} \tag{3}$$

ML-CAQR uses successive elimination trees at each recursion level r, each of which are traversed in $\log P_{r_r}$ steps. Moreover, successive trees from level l down to level r come from different recursive calls: they are inherently sequentialized. Thus, the total number of calls at a given recursion level r can be upper-bounded by $N_r = 2^{l-r} \prod_{j=r}^{l} \log P_{r_j}$. An upper bound on the global cost of *ML-CAQR* can be expressed in terms of number of calls at each level of recursion, broken down between calls performed on leaves or higher levels in the trees.

In the following γ is the flop rate and $\bar{F}_{ML\text{-}CAQR}(n,n)$ is the computational cost of *ML-CAQR* applied to a square matrix of size n. We assume that for each level k, we have $P_{r_k} = P_{c_k} = \sqrt{P_k}$, and that block sizes are chosen to make the additional costs lower order terms, that is $b_k = O(n/(\sqrt{P_k^*} \cdot \prod_{j=k}^{l} \log^2 P_j))$. Then, by expanding all recursive costs from level l down to level 1, the cost of *ML-CAQR* can be expressed as:

$$\bar{F}_{ML\text{-}CAQR}(n,n) \leq \frac{4n^3}{P}\gamma + O\left(\frac{l \cdot n^3}{P \prod_{j=1}^{l} \log P_j}\right)\gamma \tag{4}$$

$$\bar{W}_{ML\text{-}CAQR}(n,n) \leq \frac{n^2}{\sqrt{P}} \left(l \cdot \log P_1 + 4l \cdot \prod_{j=1}^{l} \log P_j + \log P_l \right) \beta_1 \tag{5}$$
$$+ \sum_{k=2}^{l-1} \frac{(l-k) \cdot n^2}{\sqrt{(P_k^*)}} \left(1 + \frac{2 \prod_{j=k}^{l} \log P_j}{\sqrt{P_l}} \right) \beta_k + \frac{n^2 \cdot \log P_l}{\sqrt{P_l^*}} \beta_l$$
$$+ O \left(\frac{l \cdot n^2}{\sqrt{P} \log P_l} \cdot \beta_1 + \sum_{k=2}^{l-1} \frac{(l-k) \cdot n^2}{\sqrt{P_k^*} \log P_l} \cdot \beta_k + \frac{n^2}{\sqrt{P_l^*} \log P_l} \cdot \beta_l \right)$$

$$\bar{S}_{ML\text{-}CAQR}(n,n) \leq l \cdot \sqrt{P} \cdot \prod_{j=1}^{l} \log^3 P_j \alpha_1 + \sum_{k=2}^{l-1} \frac{n^2 \cdot (l-k) \log P_k}{\phi_k \sqrt{P_k^*}} \alpha_k \tag{6}$$
$$+ \frac{n^2 \cdot \log P_l}{\phi_l \sqrt{P_l}} \left(1 + \frac{1}{\prod_2^{l-1} \sqrt{P_j}} \right) \alpha_l$$
$$+ O \left(\sqrt{P} \cdot \prod_{j=1}^{l} \log^2 P_j \alpha_1 + \sum_{k=2}^{l-1} \frac{(l-k) \cdot n^2}{\phi_k \sqrt{P_k^*} \log P_l} \alpha_k + \frac{n^2}{\phi_l \sqrt{P_l \log P_l}} \alpha_l \right)$$

Finally, it is important to note that the recursive nature of *ML-CAQR* can lead to three times more computations than the optimal algorithm (we ignore several lower order terms). This is similar to other recursive approaches [12]. Altogether, *ML-CAQR* allows to reach the lower bounds on communications at all levels of the hierarchy up to polylogarithmic factors. Indeed, choosing appropriate block sizes makes most of the extra computational costs lower order terms while maintaining the optimality in terms of communication. We refer the interested reader to the related research report [9] for more details on these costs.

6 Multilevel LU factorizations

Here we briefly introduce two variants of a multilevel algorithm for computing the LU factorization of a dense matrix, *ML-CALU*. Both algorithms are recursive. The first variant, 1D-*ML-CALU*, follows a uni-dimensional approach where the recursion is applied to the entire panel at each recursive call. The second variant, 2D-*ML-CALU*, processes a panel by multiple recursive calls on sub-blocks of the panel followed by a "reduction" phase similar to that of *ML-CAQR*. The base case of both recursive variants is *CALU* [13], which uses tournament pivoting to select pivot rows. 1D-*ML-CALU* has the same stability as *CALU*. However, while it minimizes bandwidth over multiple levels of parallelism, it allows to minimize latency only over one level of parallelism. 2D-*ML-CALU* which uses a two-dimensional recursive approach, is shown to be stable in practice, and reduces both bandwidth and latency over multiple levels of parallelism. A detailed description, a performance analysis, and a stability study of both algorithms can be found in [9].

We note that similar multilevel approaches can be applied in the context of the communication avoiding rank revealing QR factorization [14], as well as the communication avoiding LU factorization with panel rank revealing pivoting [15].

7 Experimental results: performance predictions

Multilevel communication avoiding algorithms are tailored for large scale platforms displaying a significant gap between processing power and communication speed. The upcoming Exascale platforms are a natural target for these algorithms. We present performance predictions on a sample exascale platform. Current petascale platforms already display a hierarchical nature which strongly impacts the performance of parallel applications. Exascale will dramatically amplify this trend. We plan to provide here an insight on what could be observed on such platforms.

Level	Type	#	Bandwidth	Latency
1	2x 6-cores Opterons	12	19.8 GB/s	1×10^{-9}s
2	Hopper nodes	2	10.4 GB/s	1×10^{-6}s
3	Gemini ASICS	9350	3.5 GB/s	1.5×10^{-6}s

Table 1. Characteristics of NERSC Hopper.

As exascale platforms are not available yet, we base our sample exascale platform on the characteristics of the NERSC Hopper [16, 17] petascale platform. It is composed of *Compute Nodes*, each with two hexacore AMD Opteron Magnycours 2.1GHz processors offering a peak performance of 8.4 GFlop/s, with 32 GB of memory. Nodes are connected in pairs to *Gemini ASICs*, which are interconnected through the *Gemini network* [18, 19]. Detailed parameters of the Hopper platform are presented in Table 1.

Level	Type	#	Bandwidth	Latency (formula)	Latency (adjusted)
1	Multi-cores	1024	300 GB/s	1×10^{-10}s	1×10^{-9}s
2	Nodes	32	150 GB/s	1×10^{-7}s	1.2×10^{-7}s
3	Interconnects	32768	50 GB/s	1.5×10^{-7}s	1×10^{-6}s

Table 2. Characteristics of a sample exascale platform.

Our target platform is obtained by increasing the number of nodes at all 3 levels, leading to a total of 1 million nodes. The amount of memory per processing element is kept constant at 1.3 GB. Moreover, exascale platforms are likely to be available around year 2018. Therefore, latencies and bandwidths are derived using an average 15% decrease per year for the latency and a 26% increase for the bandwidth [19, 18].

However, doing so might conduct to latencies so low that electrical signals would have to travel faster than the speed of light in vacuum. This is of course

impossible. Therefore, to alleviate this problem, we assume that electrical signal travels at 10% of the speed of light in copper, against 90% in fiber optics. We consider the links within a multicore processor to be made out of copper (at level 1) and the die to be at most 3cm-by-3cm. The links between a group of nodes (i.e. at level 2) are assumed to be based on fiber optics while the interconnect at the last level are assumed to be copper links. Finally, we assume the global supercomputer footprint to be 30m-by-30m. These parameters are detailed in Table 2. We model the platform with respect to the HCP model, and use it to estimate the running times of our algorithms.

We note that in order to assess the performance of multilevel algorithms, costs of state-of-the-art 1-level communication avoiding algorithms need to be expressed with respect to the HCP model. To this end, we assume (1) each communication to go through the entire hierarchy: two communicating nodes thus belong to two distant nodes of level l, hence a bandwidth β_l. (2) Bandwidth is shared among parallel communications.

We evaluate the performance of the *ML-CAQR* algorithm as well as *CAQR* on a square matrix of size $n \times n$, distributed over a square 2D grid of P_k processors at each level k of the hierarchy, $P_k = \sqrt{P_k} \times \sqrt{P_k}$. In the following, we assume all levels to be *fully-pipelined*. Similar results are obtained regarding *forward* hierarchies, which is explained by the fact that realistic test cases are not latency bound, but are mostly impacted by their bandwidth cost.

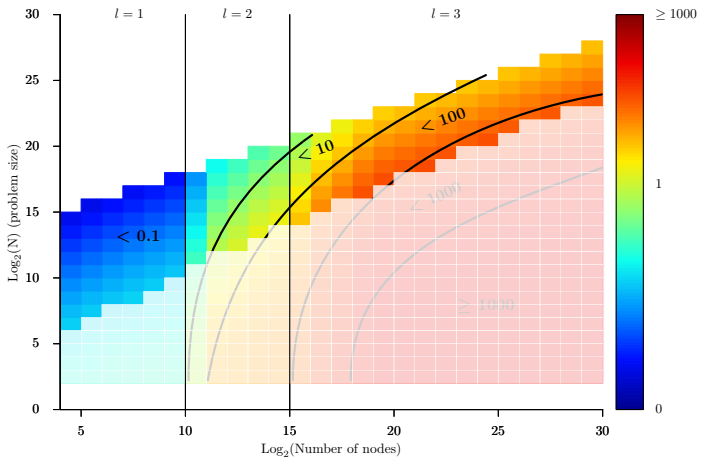


Fig. 3. Prediction of communication to computation ratio on an exascale platform for 1-level *CAQR* .

The larger the platform is, the more expensive the communication becomes. This trend can be illustrated by observing the communication to computation ratio, or *CCR* of an algorithm. In Figures 3 and 4, we plot the *CCR* of *CAQR* and *ML-CAQR* on the exascale platform. The shaded areas correspond to unrealistic cases where there are more processing elements than matrix elements and should not be considered. As the number of processing elements increases, the

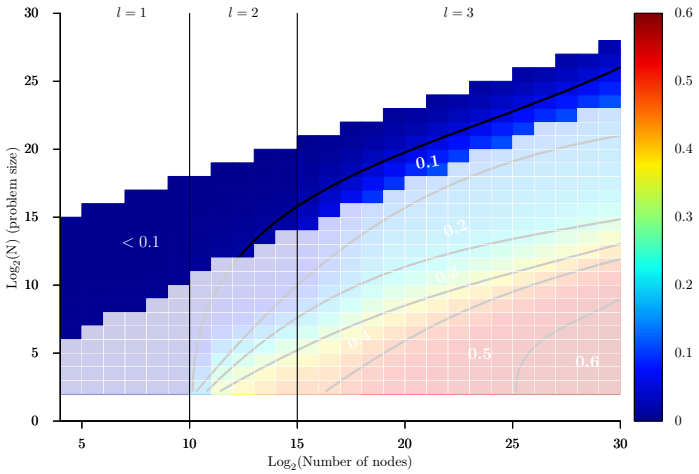


Fig. 4. Prediction of communication to computation ratio on an exascale platform for 1-level *ML-CAQR* .

cost of *CAQR* (in Figure 3) gets dominated by communication. Our multilevel approach alleviates this trend, and *ML-CAQR* (in Figure 4) allows to decrease communication, especially when the number of levels involved is large. Note that for $l = 1$, *ML-CAQR* and *CAQR* are equivalent.

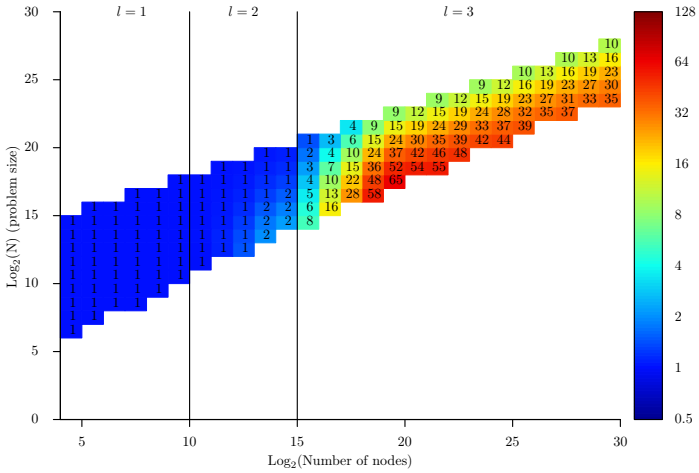


Fig. 5. Speedup of *ML-CAQR* vs. 1-level *CAQR*

However, as *ML-CAQR* performs more computations than *CAQR*, we compare the expected running times of both algorithms. Here, we denote by running time the sum of computational and communication costs. We thus assume no overlap between computation and communication. The ratio of the *ML-CAQR* running time over *CAQR* is depicted in Figure 5. *ML-CAQR* clearly outperforms *CAQR* when using the entire platform, despite its higher computational

costs. As a matter of fact in this regime, the running time is dominated by the bandwidth cost, and *ML-CAQR* significantly reduces it at all levels.

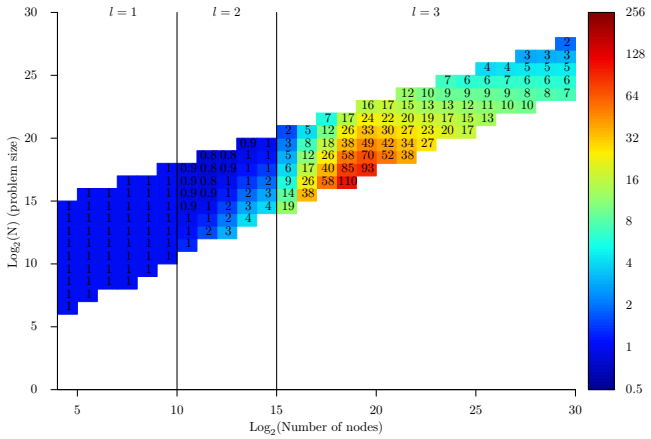


Fig. 6. Speedup of *ML-CALU* vs. 1-level *CALU*

Regarding the running times ratio, depicted in Figure 6, we can also conclude that *ML-CALU* is able to keep communication costs significantly lower than *CALU* when the entire platform is used, leading to significant speedups.

8 Conclusion

In this paper we have studied *ML-CAQR*, an algorithm that minimizes communication over multiple levels of parallelism at the cost of performing redundant computation. The complexity analysis is performed within Hcp, a model that takes into account the communication cost at each level of a hierarchical platform. The multilevel QR factorization algorithm has similar stability properties to classic algorithms. Two variants of the multilevel LU factorization have been introduced but not discussed in details. Our performance predictions on a model exascale platform show that for strong scaling, the multilevel algorithms lead to important speedups compared to algorithms minimizing communication over only one level of parallelism.

Acknowledgments. This work was supported in part by the European Research Council Advanced Grant MATFUN (267526) and the Scientific Discovery through Advanced Computing (SciDAC) program funded by U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research and Basic Energy Sciences.

References

1. Hong, J.W., Kung, H.T.: I/O complexity: The Red-Blue Pebble Game. In: STOC '81: Proceedings of the Thirteenth Annual ACM Symposium on Theory of Computing, New York, NY, USA, ACM (1981) 326–333
2. Irony, D., Toledo, S., Tiskin, A.: Communication lower bounds for distributed-memory matrix multiplication. J. Parallel Distrib. Comput. **64**(9) (2004) 1017–1026
3. Demmel, J.W., Grigori, L., Hoemmen, M., Langou, J.: Communication-optimal parallel and sequential QR and LU factorizations. SIAM Journal on Scientific Computing (2012) short version of technical report UCB/EECS-2008-89 from 2008.
4. Ballard, G., Demmel, J., Holtz, O., Schwartz, O.: Minimizing communication in numerical linear algebra. SIAM Journal on Matrix Analysis and Applications **32** (2011) 866–901
5. Agullo, E., Coti, C., Dongarra, J., Herault, T., Langou, J.: QR factorization of tall and skinny matrices in a grid computing environment. In: IPDPS'10, the 24st IEEE Int. Parallel and Distributed Processing Symposium. (2010)
6. Song, F., Ltaief, H., Hadri, B., Dongarra, J.: Scalable tile communication-avoiding QR factorization on multicore cluster systems. In: SC'10, the 2010 ACM/IEEE conference on Supercomputing, IEEE Computer Society Press (2010)
7. Bosilca, G., Bouteiller, A., Danalis, A., Faverge, M., Haidar, A., Herault, T., Kurzak, J., Langou, J., Lemarinier, P., Ltaief, H., Luszczek, P., YarKhan, A., Dongarra, J.: Flexible development of dense linear algebra algorithms on massively parallel architectures with DPLASMA. In: 12th IEEE International Workshop on Parallel and Distributed Scientific and Engineering Computing (PDSEC'11). (2011)
8. Dongarra, J., Faverge, M., Hrault, T., Jacquelin, M., Langou, J., Robert, Y.: Hierarchical QR factorization algorithms for multi-core clusters. Parallel Computing **39**(45) (2013) 212 – 232
9. Grigori, L., Jacquelin, M., Khabou, A.: Multilevel communication optimal LU and QR factorizations for hierarchical platforms. CoRR **abs/1303.5837** (2013)
10. Schreiber, R., Van Loan, C.: A storage efficient WY representation for products of Householder transformat ions. SIAM J. Sci. Stat. Comput. **10**(1) (1989) 53–57
11. Cappello, F., Fraigniaud, P., Mans, B., Rosenberg, A.: An algorithmic model for heterogeneous hyper-clusters: rationale and experience. International Journal of Foundations of Computer Science **16**(02) (2005) 195–215
12. Frens, J., Wise, D.: Qr factorization with morton-ordered quadtree matrices for memory re-use and parallelism. In: ACM SIGPLAN Notices. Volume 38., ACM (2003) 144–154
13. Grigori, L., Demmel, J., Xiang, H.: CALU: A communication optimal LU factorization algorithm. SIAM Journal on Matrix Analysis and Applications **32** (2011) 1317–1350
14. Demmel, J., Grigori, L., Gu, M., Xiang, H.: Communication avoiding rank revealing qr factorization with column pivoting. Technical Report UCB/EECS-2013-46, EECS Department, University of California, Berkeley (May 2013)
15. Khabou, A., Demmel, J., Grigori, L., Gu, M.: Lu factorization with panel rank revealing pivoting and its communication avoiding version. SIAM J. Matrix Analysis Applications **34**(3) (2013) 1401–1429
16. NERSC: Hopper configuration page. `http://www.nersc.gov/users/computational-systems/hopper/configuration`

17. Shalf, J.: Cray xe6 architecture. http://www.nersc.gov/assets/Uploads/ShalfXE6ArchitectureSM.pdf (2011)
18. Editor & lead study, P.K.: Exascale computing study: Technology challenges in achieving exascale systems (2008)
19. Graham, S., Snir, M., Patterson, C., National Research Council (U.S.). Committee on the Future of Supercomputing: Getting up to speed: the future of supercomputing. National Academies Press (2005)