A communication-optimal 2.5D LU factorization algorithm

Edgar Solomonik and James Demmel
{solomon, demmel}@cs.berkeley.edu

Communication lower bounds

The generalized communication lower bound for linear algebra [1] states that for a fast memory of size M the lower bound on communication bandwidth is

$$W = \Omega\left(\frac{\#arithmetic\ operations}{\sqrt{M}}\right)$$

words, and the lower bound on latency is

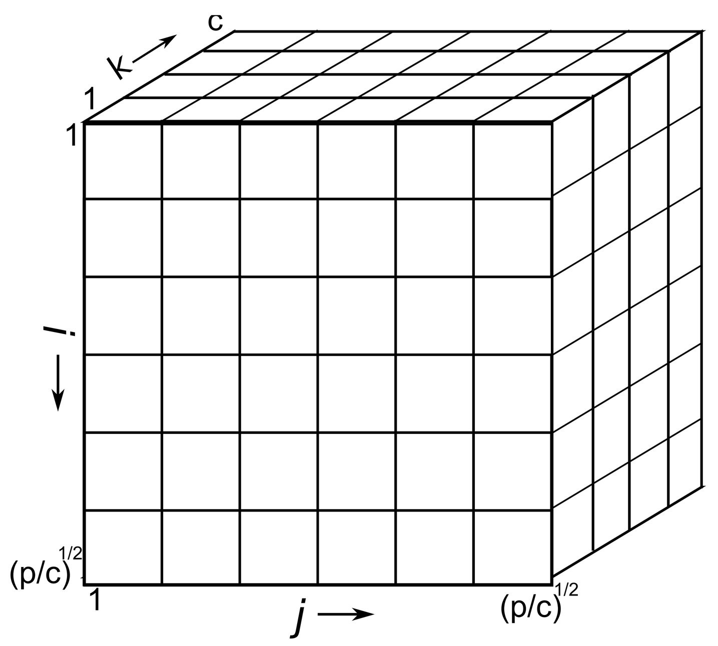
$$S = \Omega\left(\frac{\#arithmetic\ operations}{M^{3/2}}\right)$$

messages. On a parallel machine with p processors and a local processor memory of size $M = O(cn^2)$ (matrix replication factor c), this yields the following lower bounds for communication costs

$$W = \Omega\left(\frac{n^2}{\sqrt{cp}}\right),$$
$$S = \Omega\left(\sqrt{p/c^3}\right).$$

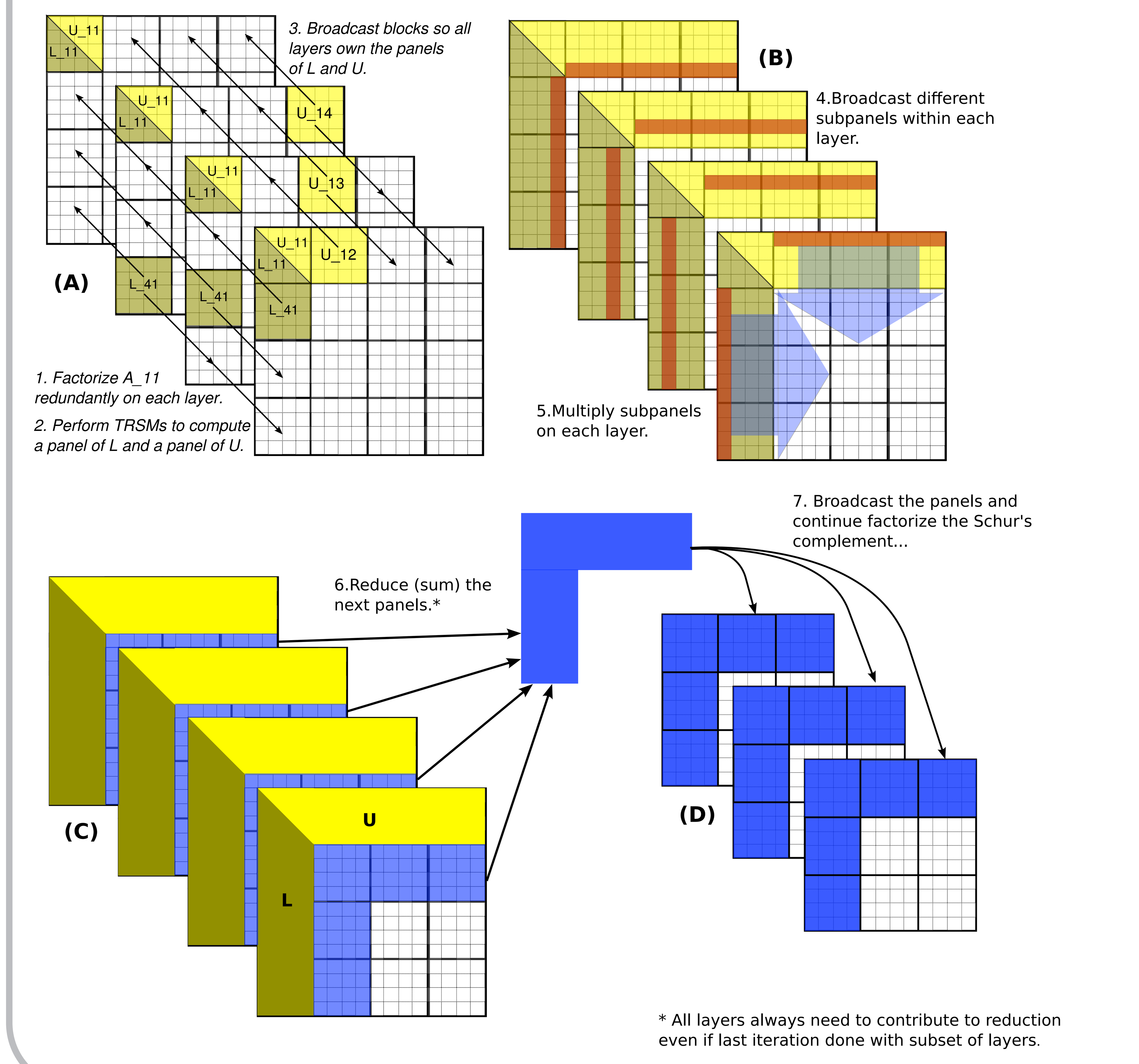
2.5D Matrix multiplication

The 2.5D processor grid is $\sqrt{p/c}$-by-$\sqrt{p/c}$-by-c as below.

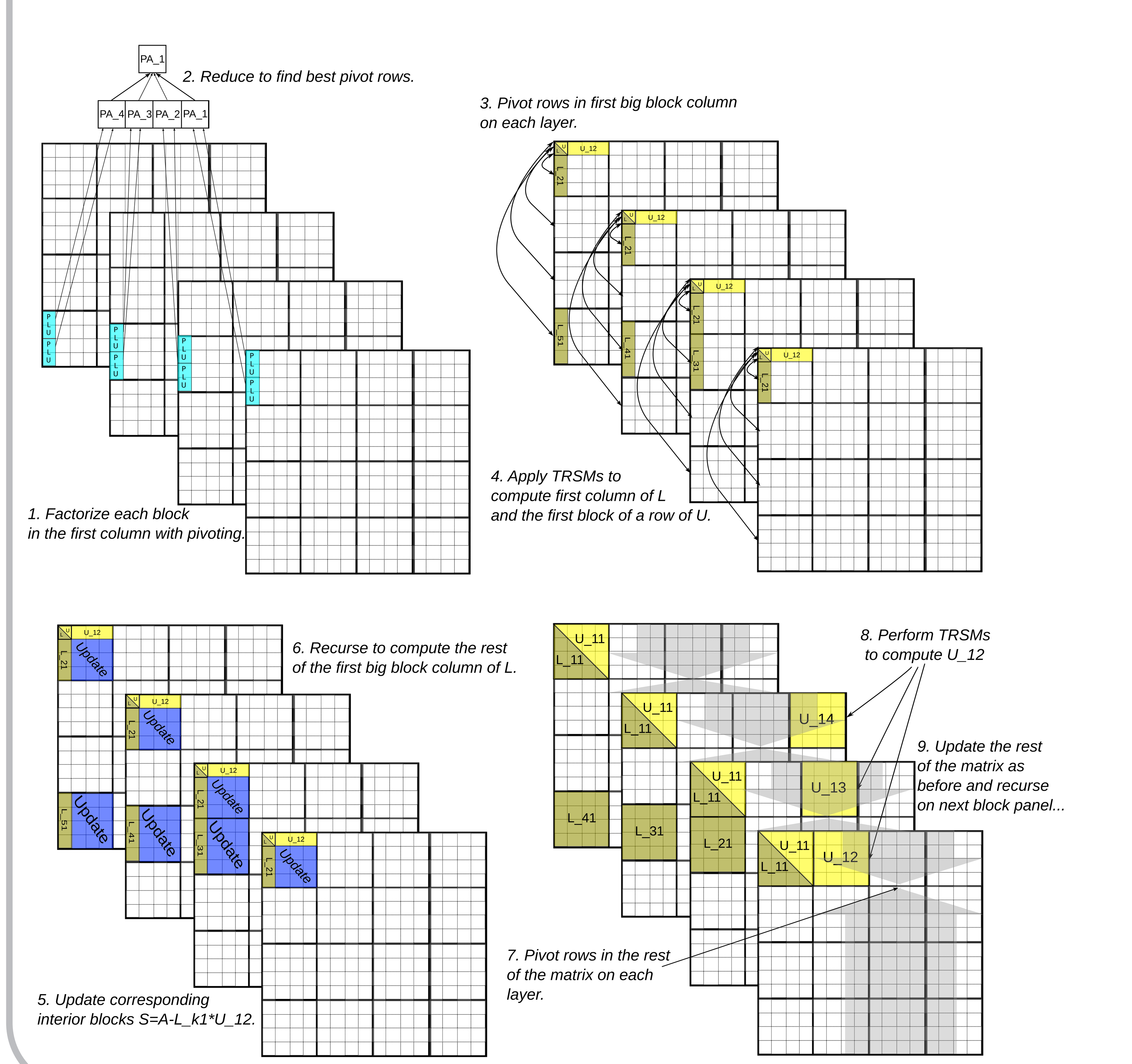


Replicate A and B over all ij and jk layers, respectively, so that P_{ijk} owns A_{ij} and B_{jk}.
Shift A rightwards by $j - i + \frac{k \cdot p^{1/2}}{c^{3/2}}$
Shift B downwards by $i - j + \frac{k \cdot p^{1/2}}{c^{3/2}}$
Multiply: $C_{ijk} = A_{local} \cdot B_{local}$
for $t = 1$ **to** $(p/c)^{1/2} - 1$ **do**
  Shift A rightwards by 1.
  Shift B downwards by 1.
  Multiple and accumulate: $C_{ijk} += A_{local} \cdot B_{local}$
end
Reduce C: $C_{ij} = \sum_{k=1}^{c} C_{ijk}$.

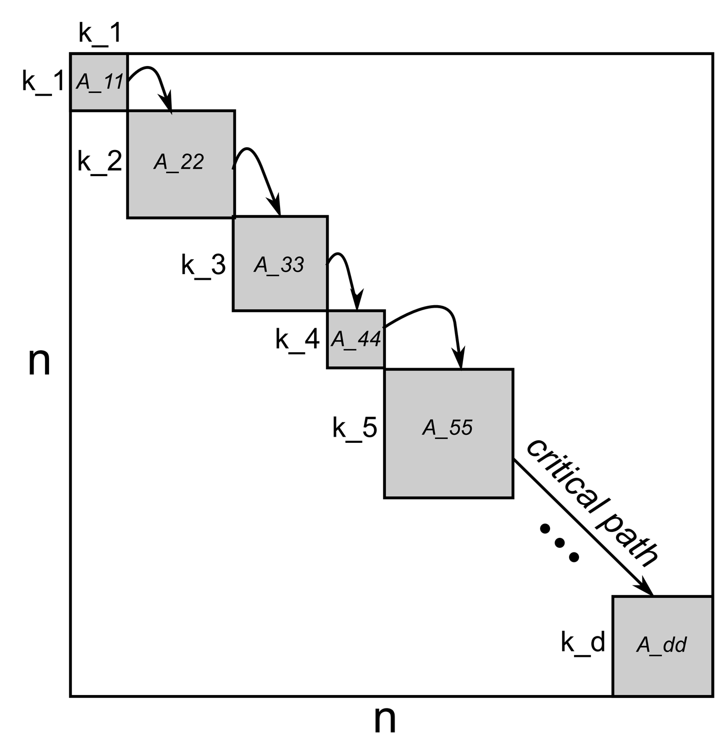
2.5D LU factorization algorithm



2.5D LU with CA-pivoting

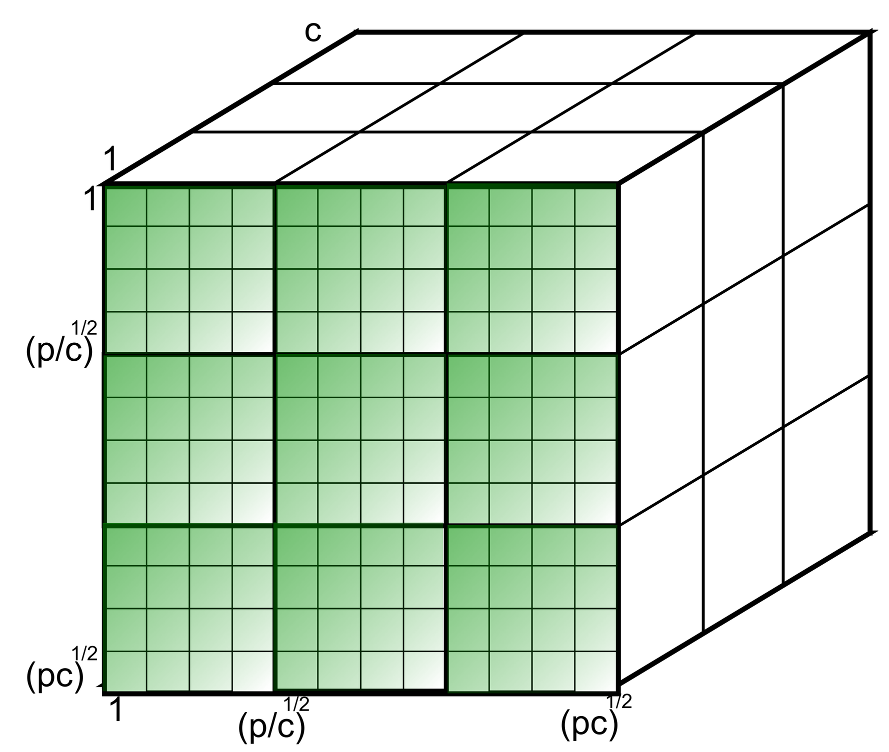


2.5D LU latency lower bound



If we measure the amount of data communicated along the critical path (1 message and k^2 words per block), we see that to achieve the bandwidth lower bound the number of messages that must be sent is at least $\Omega(\sqrt{pc})$.

2.5D LU layout



Block size of $\Omega(n/\sqrt{pc})$ is required to reach the bandwidth lower bound. So we decompose A block-cyclically on each layer. The virtualized processor grid is shown above. Each processor owns a sub-block of every big block on some layer.

Algorithm analysis

The 2.5D LU algorithm with no pivoting reaches the bandwidth lower bound and the latency lower bound within a $\log(p)$ factor. The 2.5D LU with CA-pivoting reaches all lower bounds within a $\log(p)$ factor with minor assumptions on the pivoting structure of the matrix. Both algorithms reduce to optimal 2D algorithms when $c = 1$.

References

[1] G. Ballard, J. Demmel, O. Holtz, and O. Schwartz. Minimizing communication in linear algebra. submitted to SIAM J. Mat. Anal. Appl., UCB Technical Report EECS-2009-62, 2010.

Funding

This work was supported in part by a Department of Energy CSGF fellowship, grant DE-FG02-97ER25308.