Makespan lower bound combining communication and computation for dynamic task allocation on homogeneous resource: dense Cholesky factorization test case

Mathieu Vérité
PhD thesis supervised by
Olivier Beaumont & Lionel Eyraud-Dubois
joint work with
Julien Langou, Willy Quach, Alena Shilova

INRIA Bordeaux Sud Ouest - Université de Bordeaux

SOLHARIS & HPC Scalable Ecosystems
July 2nd, 2021

1 Introduction

Test case

Cholesky factorization: $\mathbf{A} = \mathbf{L}\mathbf{L}^T$

(**A**: symmetric definite positive matrix)

Context

- **A**: $N \times N$ tiles without compression $\Rightarrow$ **homogeneous workload**
- P **homogeneous processors** with **distributed memory**
- processor to processors communication on a **shared medium** (bus)
- objective: **minimize makespan**

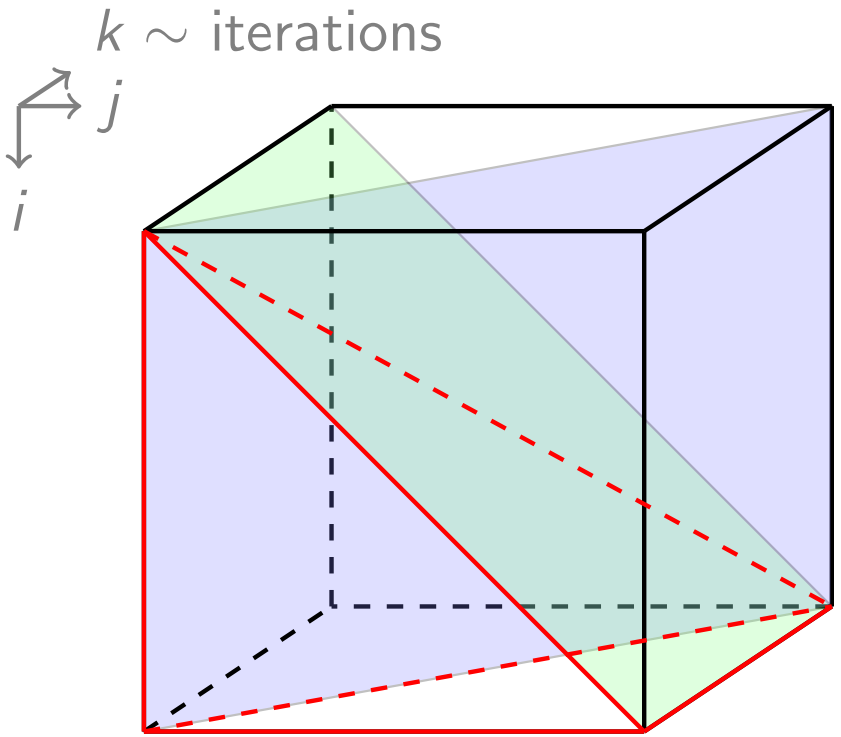
Goals

- observe **task allocation strategies** under **communication constraints**
- try to combine **computation and communication time** into a single bound

1 Introduction

Notation

- $\mathcal{T}$: set of all tasks
- w_T: computation time of task T
- c: transfer time of one tile
- $\mathcal{P}$ set of path in the DAG



3D view of the set of tasks $\mathcal{T}$

Simple bound

Assuming **no communication cost**:

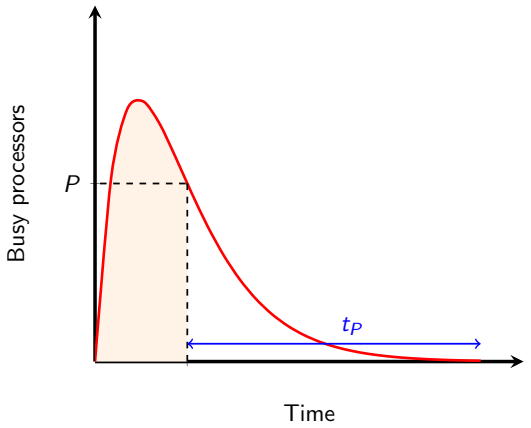
- makespan $\geqslant \frac{w_{\text{tot}}}{P} = \frac{\sum_{T \in \mathcal{T}} w_T}{P} \rightarrow$ **perfect balance**
- makespan $\geqslant \max_{p \in \mathcal{P}} \{\sum_{T \in p} w_T\} \rightarrow$ **critical path** length

2.1 Related work - *As Late As Possible* approach

Assuming

- no communication cost
- no resource limitation

$\Rightarrow$ ALAP schedule is **optimal** for minimizing the makespan

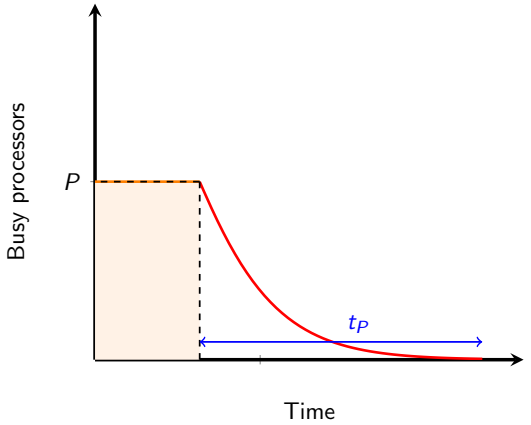


ALAP schedule: unlimited resources

Beaumont et al. (2020) [2]

From an optimal ALAP schedule: get **number of tasks in progress** at any time.

$\Rightarrow$ **makespan lower bound**



ALAP schedule: P resources

2.2 Related work - Communication lower bound

Model: each processor $\leftrightarrow$ limited, local memory of size M

Ballard et al. (2010) [1]

Asymptotic bound: $\#\mathrm{I/O} \geqslant \Omega(\frac{N^3}{\sqrt{M}})$

Olivry et al. (2019) [5]

Bound for sequential case: $\#\mathrm{I/O} \geqslant \frac{1}{6}\frac{N^3}{\sqrt{M}}$

Kwasniewski et al. (2020) [3]

Parallel case for LU factorization

- bound: $\#\mathrm{I/O} \geqslant \frac{2}{3}\frac{N^3}{\sqrt{M}}$
- algorithm: $\#\mathrm{I/O} = \frac{N^3}{\sqrt{M}}$

For Cholesky can expect:

$$\#\mathrm{I/O} \geqslant \frac{1}{3}\frac{N^3}{\sqrt{M}}$$

3.1 Contribution - Loomis-Whitney like approach

Problem (3D)

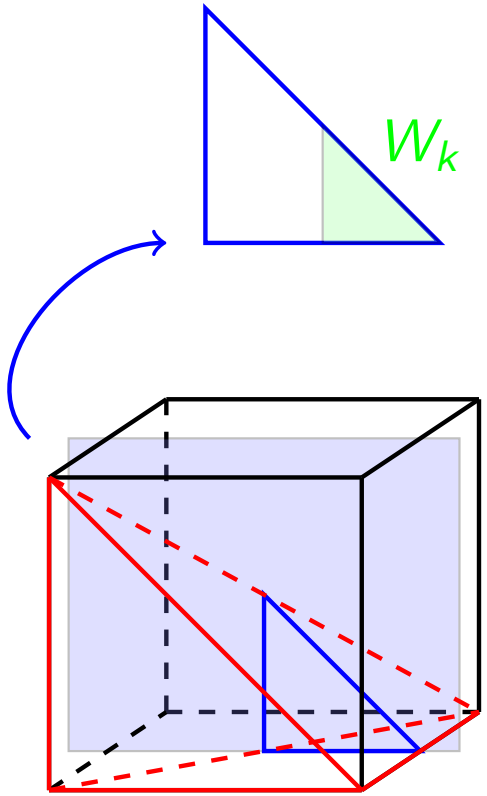
Let W tasks $\rightarrow$ which $\mathcal{W} \subset \mathcal{T}$ of size W minimizes #I/O?

$$\min\{\#\text{I/O required for } \mathcal{W}\} \text{ such that } |\mathcal{W}| = W$$

Resolution steps

1. iteration k (2D problem): optimal solution $W_k \rightarrow \approx \sqrt{2W_k}$ #I/O
2. problem becomes:

$$\min\{\sum_k \sqrt{2W_k} + \max\{W_k\}\}$$
$$\text{st.} \begin{cases} W_k \leqslant \max\{W_k\} \\ \sum_k W_k = W \end{cases}$$



W tasks requires at least $\frac{3}{2^{\frac{1}{3}}} W^{\frac{2}{3}}$ #I/O.

3.2 Contribution - Operational intensity & communication

Operational Intensity (ρ)

$$\rho = \frac{\#\text{computation}}{\#\text{I/O}}$$

Minimum data movement
$\Leftrightarrow$ maximum ρ

1. W tasks $\rightarrow$ $\#\text{I/O} \geqslant \frac{3}{2^{\frac{1}{3}}} W^{\frac{2}{3}}$
2. M $\#\text{I/O} \rightarrow \#\text{computation} \leqslant \frac{\sqrt{2}}{3\sqrt{3}} M^{\frac{3}{2}}$

$$\rho \leqslant \frac{\sqrt{2}}{3\sqrt{3}} \sqrt{M}$$

Communication bound

Assumptions

- **perfect load balancing**
- only one copy of **A** fits in memory: $M = \mathcal{O}(\frac{N^2}{P}) \rightarrow$ using all P

Bound

$\approx \frac{N^3}{6}$ tasks in Cholesky

$$\#\text{I/O} \geqslant \frac{\sqrt{3}}{2\sqrt{2}} \frac{N^3}{\sqrt{M}} \approx 0.61 \frac{N^3}{\sqrt{M}} > \frac{1}{3} \frac{N^3}{\sqrt{M}}$$

3.3 Contribution - Combined bound

Communication bound assumes **use all** P.
But **opposite effects**:

- use **few** $P \Rightarrow \searrow$ #I/O
- use **many** $P \Rightarrow \nearrow$ parallelism

Optimal number of P to use?

Mix computation and communication

1. Let $\gamma > 1$ and L_p processor p workload. Assume:
$$\forall p \in \{1, \ldots, P\},\ L_p \leqslant \gamma \frac{w_{\text{tot}}}{P} = w_{\text{max}}$$
2. $\min\{\#\text{I/O}\} \rightarrow L_1, \ldots, L_P = w_{\text{max}}, \ldots, w_{\text{max}}, 0, \ldots, 0$
$$\Rightarrow t_{\text{comm}} \geqslant c(\frac{P}{\gamma} - 1)\frac{3}{2^{\frac{1}{3}}}(\gamma \frac{w_{\text{tot}}}{P})^{\frac{2}{3}} \text{ (and } t_{\text{compute}} = \gamma \frac{w_{\text{tot}}}{P})$$
3. Optimal situation $t_{\text{compute}} = t_{\text{comm}}$
$$\Rightarrow \frac{P}{\gamma} \leqslant 1 + \frac{1}{3}(\frac{w}{c}N)^{\frac{3}{4}} \quad (\text{for } w_{\text{tot}} \approx \frac{N^3}{6})$$

Combined bound for speed-up: $1 + \frac{1}{3}(\frac{w}{c}N)^{\frac{3}{4}}$

4.1 Experiments - Setup

Testing parameters

- number of resource P
- bus bandwidth
- tasks allocation strategies

$\Rightarrow$ **Simulation** of execution on a **single node**.

Simgrid

- topology: **single shared bus**
- **no latency** cost

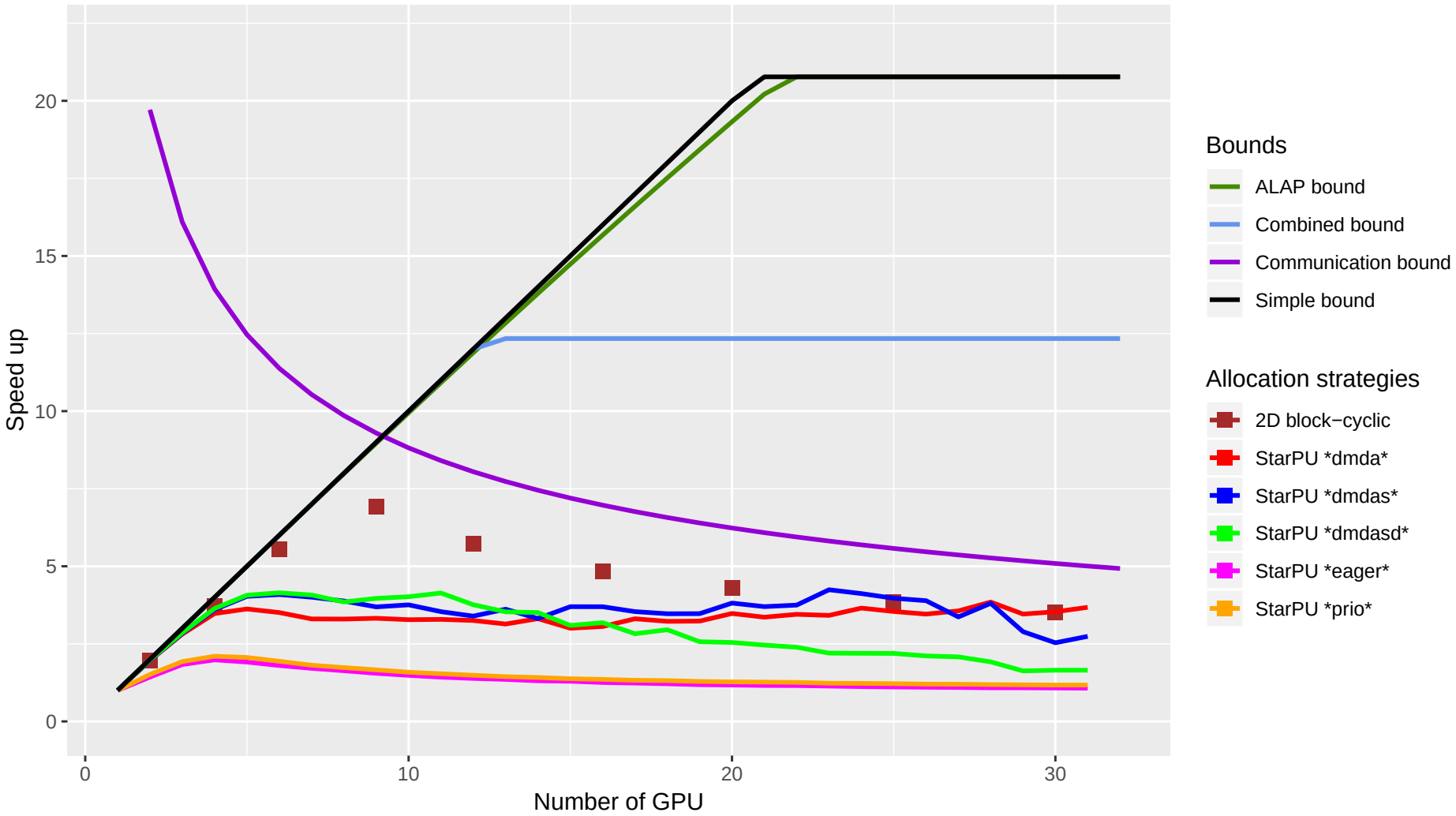
StarPU

- **deterministic performance model**
- no cost for "handling" tasks

Chameleon

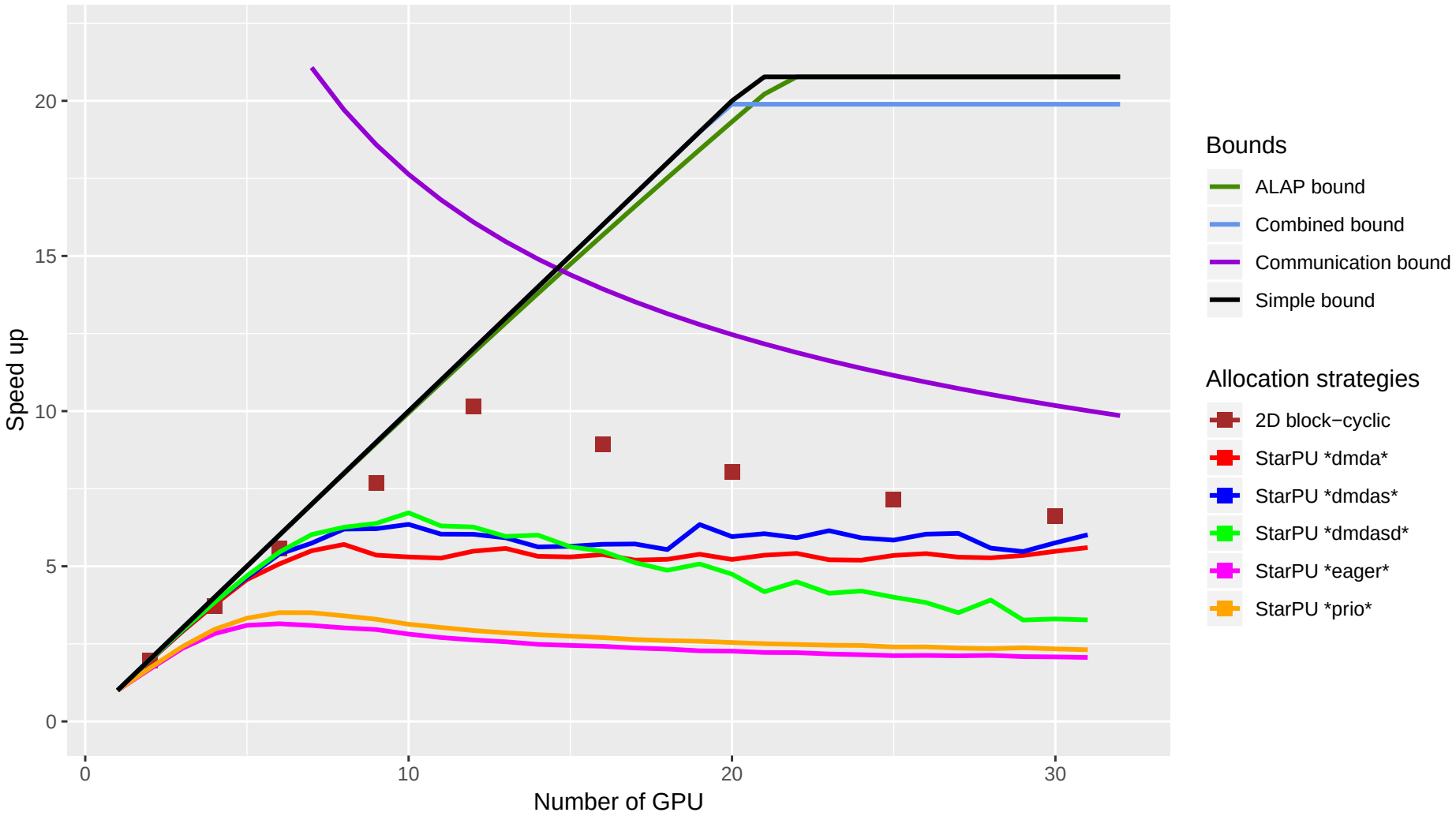
- tasks priorities: discrete levels

4.2 Experiments - Some results



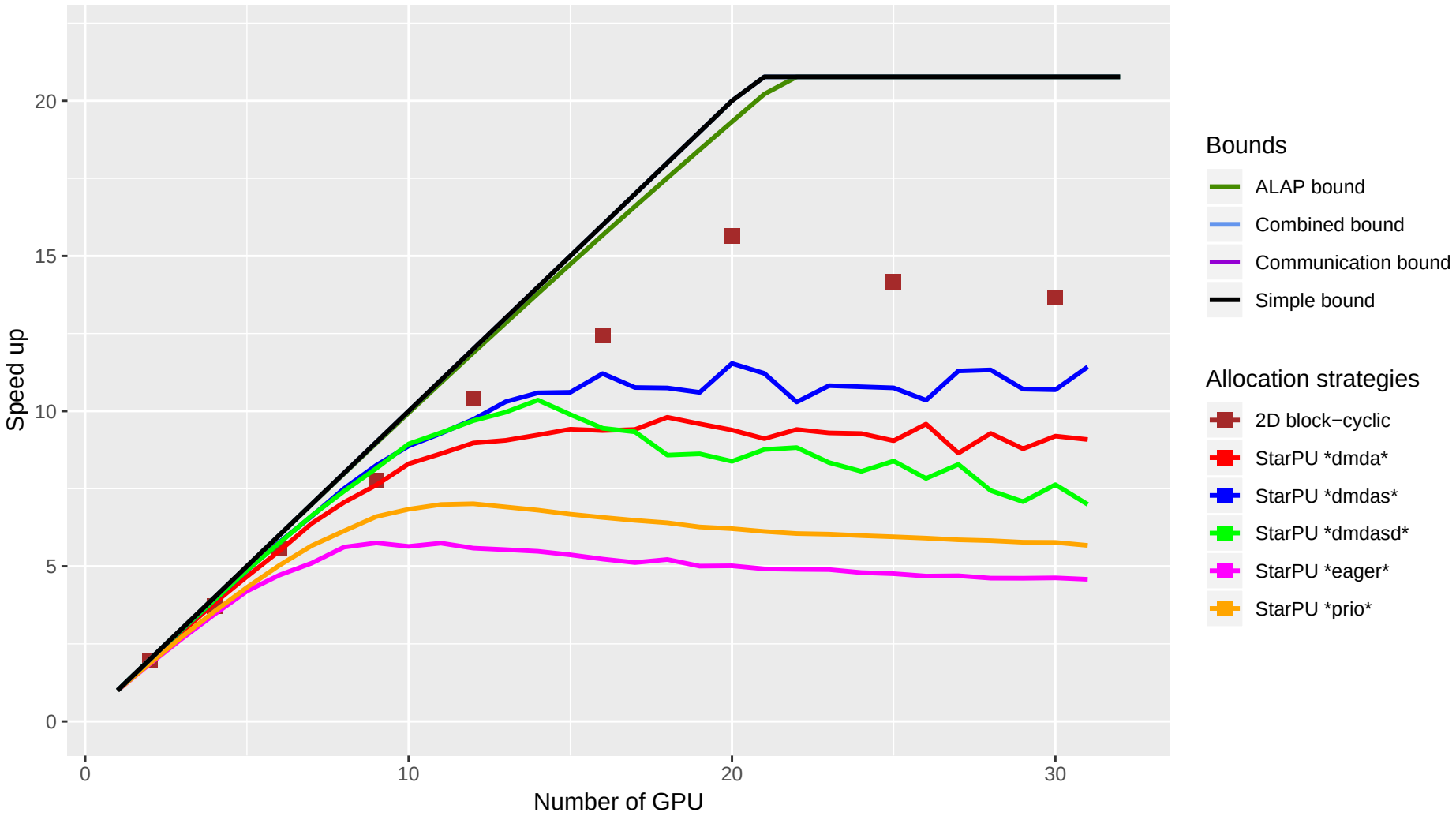
30×30 Cholesky factorization - 10GB/sec bandwidth

4.2 Experiments - Some results



30×30 Cholesky factorization - 20GB/sec bandwidth

4.2 Experiments - Some results



30×30 Cholesky factorization - 50GB/sec bandwidth

4.3 Experiments - Observations

Allocation strategies

- evolution of performance **consistent** with communication constraint
- *2D block-cyclic* **close to communication bound**
- *dmda[s]* maintains good performance even for large P

Bounds

- new communication bound **relevant**
- combined computation/communication bound: "optimal" P **shifted from performance peak**

5 Conclusion

Results

- communication lower bound:
  - improvement from state-of-the art (0.61 VS $\frac{1}{3}$)
  - how tight? (there is a solution with $\#I/O = \frac{N^3}{\sqrt{M}}$)
- combined bound $\rightarrow$ hint for optimal P to use
  but limited to bus topology

Future work

- communication bound getting tighter $\rightarrow$ refine geometric method
- extension to cases using CPU and GPU

Reference


Ballard, G., Demmel, J., Holtz, O., and Schwartz, O.
Communication-optimal parallel and sequential cholesky decomposition.
CoRR abs/0902.2537 (2009).


Beaumont, O., Langou, J., Quach, W., and Shilova, A.
A Makespan Lower Bound for the Scheduling of the Tiled Cholesky Factorization based on ALAP Schedule.
In *EuroPar 2020 - 26th International European Conference on Parallel and Distributed Computing* (Warsaw / Virtual, Poland, Aug. 2020), Proceedings of EuroPar 2020, Springer.


Kwasniewski, G., Ben-Nun, T., Ziogas, A. N., Schneider, T., Besta, M., and Hoefler, T.
On the parallel I/O optimality of linear algebra kernels: Near-optimal LU factorization.
CoRR abs/2010.05975 (2020).


Loomis, L. H., and Whitney, H.
An inequality related to the isoperimetric inequality.
Bulletin of the American Mathematical Society 55, 10 (1949), 961 – 962.


Olivry, A., Langou, J., Pouchet, L., Sadayappan, P., and Rastello, F.
Automated derivation of parametric data movement lower bounds for affine programs.
CoRR abs/1911.06664 (2019).

Thank you for your attention