Communication Bounds for Heterogeneous Architectures

Grey Ballard, James Demmel, Andrew Gearhart

{ballard,demmel,agearh}@cs.berkeley.edu

Summary

- New communication lower bounds for nearly all direct linear algebra problems on heterogeneous architectures
- New algorithms that attain lower bounds
- Preliminary empirical results that support theory

Motivation/Background

Communication

- Defined as data movement between processors and global memory
- Measured as # words (inverse bandwidth) and # messages (latency)
- Matters because it's much slower relative to flops...and this is getting worse

Established Communication Bounds

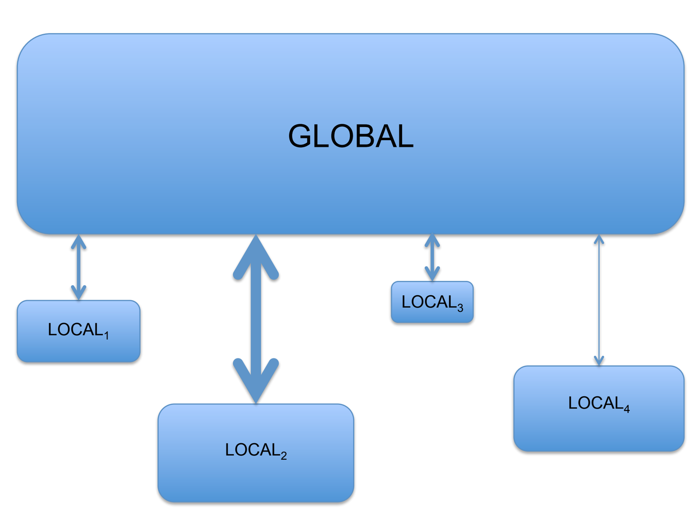
Lower bound on:	Lower bound
# words (W)	max $\left(\#\text{inputs} + \#\text{outputs}, \#\text{flops} / (\text{fast memory size})^{1/2}\right)$
# messages (L)	max $\left(\#\text{inputs} + \#\text{outputs}, \#\text{flops} / (\text{fast memory size})^{3/2}\right)$

- Results due to Ballard/Demmel/Holtz/Schwartz [BDHS09], Hong/Kung [HK81], Irony/Tishkin/Toledo [ITT04]

Model

Outline

- Consider a heterogeneous machine to be a collection of P compute elements linked via a global memory
- We assume that the problem data initially lives in global memory and allow each proc$_i$ to be described according several machine parameters



Machine Parameters

- M_i: Size of the local memory of proc$_i$
- γ_i: Floating point performance of proc$_i$ (seconds/flop)
- β_i: Inverse bandwidth of proc$_i$ (seconds/word)
- α_i: Latency of proc$_i$ (seconds/message)

Lower Bounds

- Time cost of message with w words: $T_{msg} = \alpha + \beta w$
- proc$_i$'s runtime: $T_i = \gamma_i F_i + \beta_i W_i + \alpha_i L_i$
- General bound on parallel runtime (I = #inputs, O = #outputs, G = total flops):

$$T \geq \min_{\sum F_i = G} \max_{1 \leq i \leq P} \gamma_i F_i + \beta_i \max\left\{I_i + O_i, \frac{F_i}{8\sqrt{M_i}}\right\} + \alpha_i \max\left\{\frac{I_i + O_i}{M_i}, \frac{F_i}{8M_i^{3/2}}\right\}$$

  – See [BDG11] for details and proof

BLAS2-type bound

- Let $\xi_i = \gamma_i + \beta_i + \frac{\alpha_i}{M_i}$
- We obtain $T \geq \max_{1 \leq i \leq P} \xi_i F_i = \frac{G}{\sum \frac{1}{\xi_j}}$

  where

$$F_i = \frac{\frac{1}{\xi_i}}{\sum \frac{1}{\xi_j}} G \qquad (1)$$

BLAS3-type bound

- Let $\delta_i = \gamma_i + \frac{\beta_i}{8\sqrt{M_i}} + \frac{\alpha_i}{8M_i^{3/2}}$
- We obtain $T \geq \max_{1 \leq i \leq P} \delta_i F_i = \frac{G}{\sum \frac{1}{\delta_j}}$

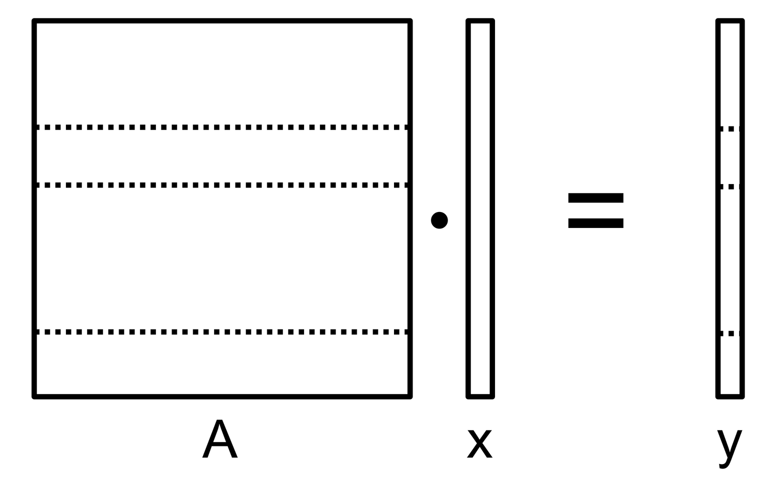
  where

$$F_i = \frac{\frac{1}{\delta_i}}{\sum \frac{1}{\delta_j}} G \qquad (2)$$

New Algorithms

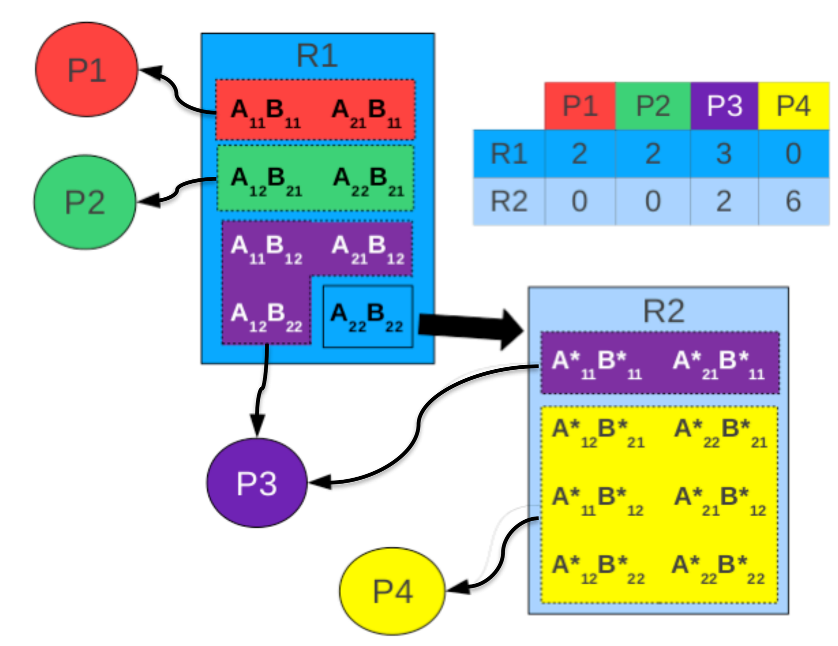
Heterogeneous Matrix-Vector Multiplication (HGEMV)

- Assume input matrix is stored in row-major format
- Set flop distribution according to Equation (1)
- Split matrix row-wise
- Each processor computes its portion of the result



Heterogeneous Matrix-Matrix Multiplication (HGEMM)

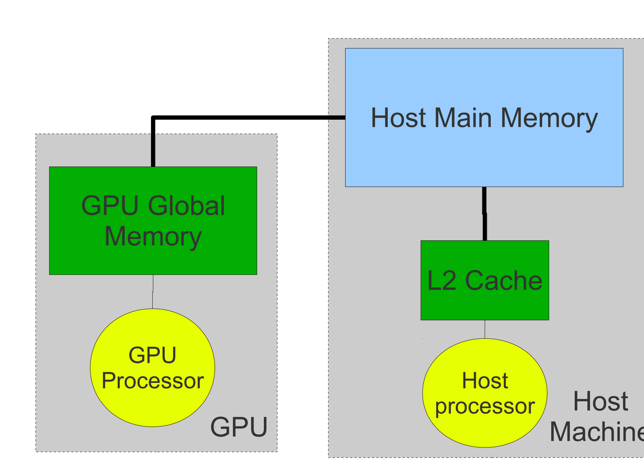
- Assume input matrix is stored in a block-recursive format
- Set flop distribution according to Equation (2)
- Convert each fraction of flops to octal: $0.d_1^{(i)} d_2^{(i)} \cdots d_k^{(i)}$
- Using square recursive GEMM, assign $d_j^{(i)}$ subproblems at level j of the recursion to proc$_i$
- Each processor computes its assigned subproblems using square recursive GEMM

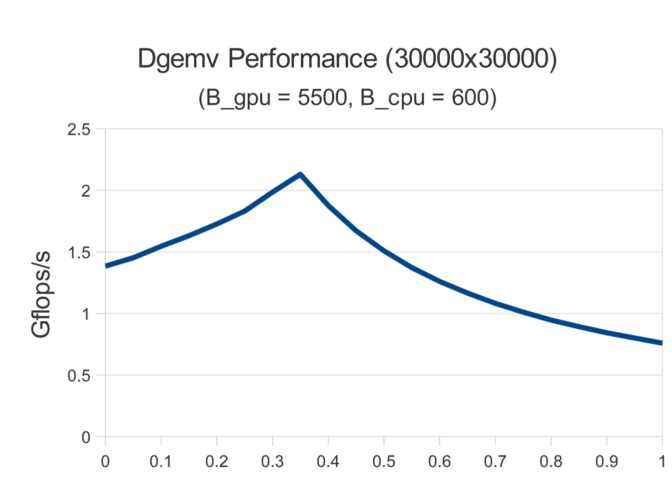


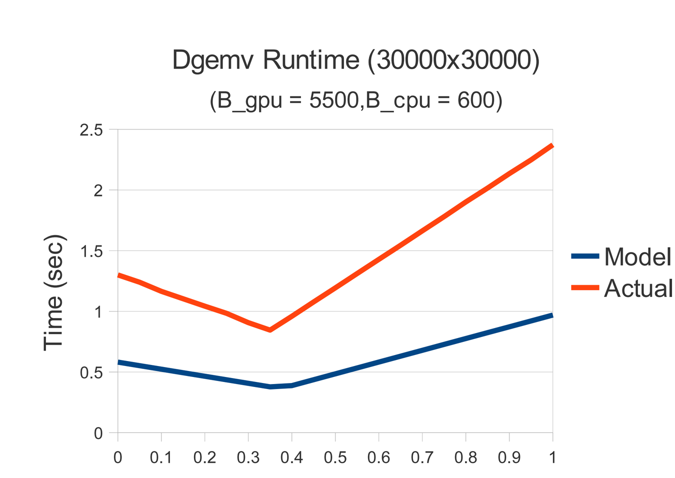
Preliminary Results

Heterogeneous Matrix-Vector Multiplication (HGEMV)

- CPU/GPU System (Intel Xeon E5405 CPU and GTX280 GPU)
- host DRAM was considered to be "global memory"
- only one core of the CPU was used for results
- Runtime bound accurately predicted optimal work distribution







Credits

- Research supported by Microsoft (Award #024263) and Intel (Award #024894) funding and by matching funding by U.C. Discovery (Award #DIG07-10227).

References

[BDG11] G. Ballard, J. Demmel, and A. Gearhart. Communication lower bounds for heterogeneous architectures, 2011. Submitted to ACM SPAA.

[BDHS09] G. Ballard, J. Demmel, O. Holtz, and O. Schwartz. A general communication lower bound for linear algebra, 2009. Submitted to SIMAX.

[HK81] Jia-Wei Hong and H. T. Kung. I/O complexity: The red-blue pebble game. In *STOC '81: Proceedings of the thirteenth annual ACM symposium on theory of computing*, pages 326–333, New York, NY, USA, 1981. ACM.

[ITT04] D. Irony, S. Toledo, and A. Tiskin. Communication lower bounds for distributed-memory matrix multiplication. *J. Parallel Distrib. Comput.*, 64(9):1017–1026, 2004.