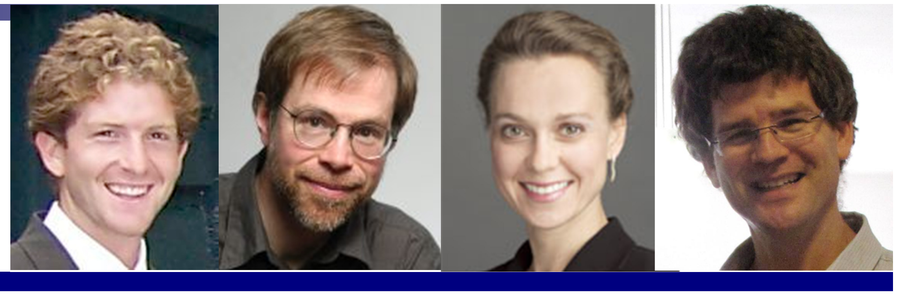
Graph Expansion and Communication Costs of Fast Matrix Multiplication

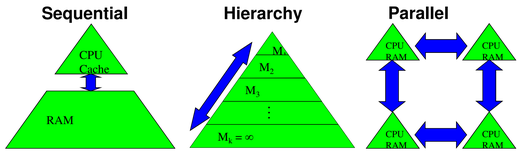
Berkeley Parlab



Grey Ballard Jim Demmel Olga Holtz Oded Schwartz

Communication-cost is Graph-expansion

Motivation



Two kinds of costs:

- Arithmetic (FLOPs)
- Communication: moving data between
  - levels of a memory hierarchy (sequential)
  - over a network connecting processors (parallel)

Communication-efficient algorithm:

Save **time**, save **energy**.

Results: New Lower bounds

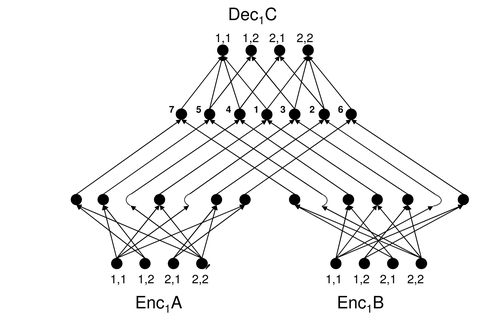
For Strassen's:	Strassen-like:	Recall for cubic:
$\Omega\left(\left(\frac{n}{\sqrt{M}}\right)^{\log_2 7} M\right)$	$\Omega\left(\left(\frac{n}{\sqrt{M}}\right)^{\omega_0} M\right)$	$\Omega\left(\left(\frac{n}{\sqrt{M}}\right)^{\log_2 8} M\right)$
$\Omega\left(\left(\frac{n}{\sqrt{M}}\right)^{\log_2 7} \frac{M}{P}\right)$	$\Omega\left(\left(\frac{n}{\sqrt{M}}\right)^{\omega_0} \frac{M}{P}\right)$	$\Omega\left(\left(\frac{n}{\sqrt{M}}\right)^{\log_2 8} \frac{M}{P}\right)$

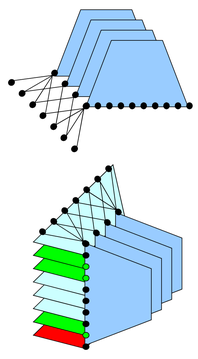
The parallel lower bounds applies to

2D: $M = \Theta(n^2/P)$

2.5D: $M = \Theta(c \cdot n^2/P)$

The CDAG of Strassen

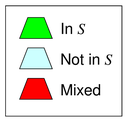




Recursive construction

Given Dec_iC, Construct $Dec_{i+1}C$:

1. Duplicate 4 times
2. Connect with a cross-layer of Dec_1C



Strassen's Fast Matrix Multiplication

Compute 2 x 2 matrix multiplication using 7 multiplications (instead of 8).

$M_1 = (A_{11} + A_{22}) \cdot (B_{11} + B_{22})$
$M_2 = (A_{21} + A_{22}) \cdot B_{11}$
$M_3 = A_{11} \cdot (B_{12} - B_{22})$
$M_4 = A_{22} \cdot (B_{21} - B_{11})$
$M_5 = (A_{11} + A_{12}) \cdot B_{22}$
$M_6 = (A_{21} - A_{11}) \cdot (B_{11} + B_{12})$
$M_7 = (A_{12} - A_{22}) \cdot (B_{21} + B_{22})$

$C_{11} = M_1 + M_4 - M_5 + M_7$
$C_{12} = M_3 + M_5$
$C_{21} = M_2 + M_4$
$C_{22} = M_1 - M_2 + M_3 + M_6$

Apply recursively (block-wise)

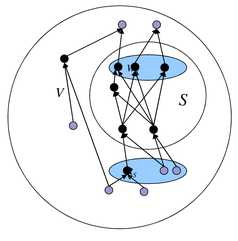
$$\begin{matrix} n/2 \\ n/2 \end{matrix}\left\{\begin{array}{|c|c|}\hline C_{11} & C_{12} \\ \hline C_{21} & C_{22} \\ \hline \end{array}\right. = \begin{array}{|c|c|}\hline A_{11} & A_{12} \\ \hline A_{21} & A_{22} \\ \hline \end{array} \cdot \begin{array}{|c|c|}\hline B_{11} & B_{12} \\ \hline B_{21} & B_{22} \\ \hline \end{array}$$

$T(n) = 7 \cdot T(n/2) + O(n^2)$

$T(n) = \Theta(n^{\log_2 7})$

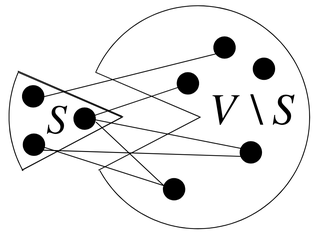
Novel Approach: Expansion

The Computation Directed Acyclic Graph



- Input / Output
- Intermediate value
- Dependency

Edge Expansion



$G = (V,E)$ is a d-regular graph with

$$h \equiv \min_{S,|S| \le \frac{|V|}{2}} \frac{|E(S,\bar{S})|}{d|S|}$$

eigenvalues: $1 = \lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_n$

$\gamma \equiv 1 - max\{\lambda_2, |\lambda_n|\}$

Thm: [Alon-Milman84, Dodziuk84, Alon86]

$$\frac{1}{2}\gamma \le h \le \sqrt{2\gamma}$$

SPAA 2011