Parallel algorithms for Dense Matrix Problems

Chris Kauffman

Last Updated:
Mon Oct 4 09:43:36 AM CDT 2021

Logistics

Assignments

- A1 grading delayed due to personal emergency for TA James, complete later this week
- Then Mini-Exam 1 grading will commence
- Goal to have all grading done by Mon 10/11
- A2 is delayed: apologies

Reading

Grama Ch 8 on Dense Matrix Algorithms

- Naive Matrix Multiply
- Cannon's Algorithm
- LU Decomposition

Today

Matrix algorithms

Recall Matrix Transpose

- Common operation on matrices is a **transpose** notated A^T
- Interchanges rows/columns of A: $a_{ij} \rightarrow a_{ji}$
- Diagonal elements stay the same
- Algorithms that perform operations on A can often be performed on A^T without re-arranging A - how? *Hint: consider summing rows of A vs summing rows of A^T*

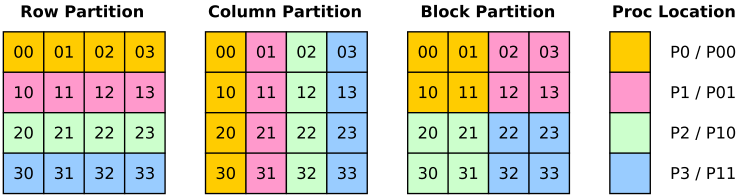
Original matrix A

0	5	10	15
20	25	30	35
40	45	50	55
60	65	70	75

transpose(A)

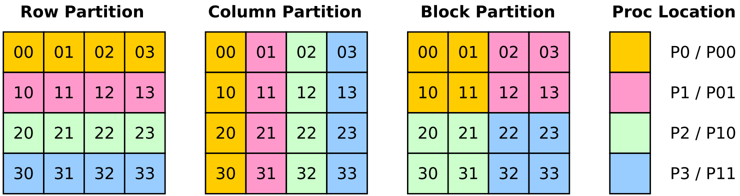
0	20	40	60
5	25	45	65
10	30	50	70
15	35	55	75

Exercise: Matrix Partitioning Across Processors



- Recall several ways to partition matrices across processors
- Diagram shows these
  - Entry `ij` may be an individual element OR...
  - Entry `ij` may be a **Block**: ex. Block (2,3) is the submatrix from rows 200-299 and cols 300-399
- Assume **square** matrices : #rows = #cols
- Common to multiply to compute product: $C = A \times B$
- Ideal partitioning for A and B in matrix multiply?
- Ideal partitioning for $C = A^T \times B$
- Ideal partitioning for $C = A \times B^T$

Answers: Matrix Partitioning Across Processors



- $C = A \times B$
  - Ideally A is row-partitioned, B is column partitioned
  - Then block-partitioned C could be computed w/o communication
- $C = A^T \times B$
  - Ideally A and B column-partitioned
- $C = A \times B^T$
  - Ideally A and B row-partitioned
- Block-partitioning often used: not ideal for any version but less communication required when both A and A^T will b used

Naive Parallel Dense Multiplication: Overview

Block Partitioning Appears Frequently

- Specific applications may be able to select a favorable partitioning (e.g. Row Partition for Page Rank)
- Many applications use both A and A^T so employ block-partitioned matrices: middle-way approach which does not favor rows or columns
- Parallel Libraries often use block partitions by default

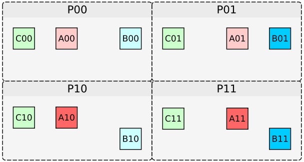
Matrix Multiply with Blocks

- To compute Matrix-Matrix multiply, procs must (eventually) multiply full rows by full columns to compute an output block
- Naive method: each Proc stores full rows/columns needed for it to independently compute output block which it stores

Naive Parallel Dense Multiplication: Demo

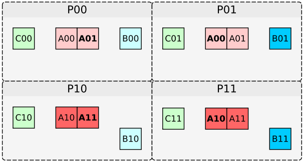
1. Initial data layout: each Proc holds a block of A, B, and C respectively. Processors are arragned in a logical grid that reflects their initial data.

In subsequent steps, received / computed data is bolded.



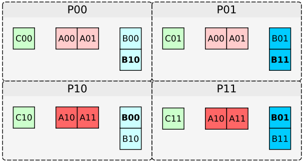
2. Each proc participates in an All-to-All sharing of data for the Row it is in.

This leaves each proc with entire rows of A.

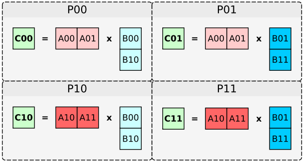


3. Each proc participates in an All-to-All sharing of data for the Column it is in.

This leaves each row with complete columns as well.



4. Each Proc now has a unique set of complete Rows and Columns and can independently compute a block of output matrix C through block multiplication.



Exercise: Analysis of Naive Dense Mult.

Assumptions

- Matrices A and B are size $N \times N$ so N^2 elements
- P processors with block partitioning: initially $N^2/\sqrt{P}$ elements of A, B on each proc (assume P is a perfect square)
- Simplified communication cost for All-to-All on a Ring:

$$t_{comm} = (p-1)(t_s + t_w M)$$

with p #procs in ring, t_s comm startup time, t_w per word transfer rate, M message size.

Questions

1. What is communication cost of this algorithm?
2. How much time does the final block matrix multiply take?
3. What is the memory requirement for each proc?
4. Biggest disadvantage for this algorithm?

Answers: Analysis of Naive Dense Mult.

1. What is communication cost of this algorithm?
   - #Procs in rows/cols is $\sqrt{P} \sim$ ring size
   - $M = N^2/\sqrt{P}$: message size is num elements on each proc
   - 2 All-to-All shares : 1 for rows, 1 for cols

   $$t_{comm} = 2(\sqrt{P} - 1) \times (t_s + t_w(N^2/\sqrt{P}))$$

2. What is the memory requirement for each proc? E.g. how many submatrices of A,B are on each proc?
   - Full rows/cols on each proc
   - Requires $\sqrt{P}$ submatrices for each Proc
3. How much time does the final block matrix multiply take?
   - Each proc has $\sqrt{P}$ submats of A,B to multiply
   - Each submat is size $N/\sqrt{O}$ with size s requiring $O(s^3)$ opts

   $$t_{mult} = O((\sqrt{P}) \times ((N/\sqrt{P})^3)) = O(N^3/P)$$

4. Biggest disadvantage for this algorithm?
   - Major: The need to store $\sqrt{P}$ sub matrices on all procs may be prohibitive
   - Minor: Not much chance to overlap communication / computation in the algorithm

Cannon's Algorithm

- Proposed in Lynn Elliot Carter's 1969 thesis
- Target was very small parallel machines implementing a Kalman Filter algorithm in hardware
- "Communication" happening between small Procs with data in registers
- Scales nicely to large machines and overcomes the large memory requirement of the Naive Mat-Mult Algorithm

A CELLULAR COMPUTER TO IMPLEMENT

THE KALMAN FILTER ALGORITHM

by

LYNN ELLIOT CANNON

By the conventional definition of matrix product, if A is multiplied by B, the result, call it C, is given by

$$C = AxB = \begin{bmatrix} a_1b_1+a_2b_2+a_3b_3 & a_1b_4+a_2b_5+a_3b_6 & a_1b_7+a_2b_8+a_3b_9 \\ a_4b_1+a_5b_2+a_6b_3 & a_4b_4+a_5b_5+a_6b_6 & a_4b_7+a_5b_8+a_6b_9 \\ a_7b_1+a_8b_2+a_9b_3 & a_7b_4+a_8b_5+a_9b_6 & a_7b_7+a_8b_8+a_9b_9 \end{bmatrix}.$$

The symmetry of this product can be seen by comparing the ij^{th} element with the ji^{th} element and noticing that one is obtained from

-24-

A. 1. The first row of A is left alone.

2. The second row of A is shifted left one column.

3. The third row of A is shifted left two columns.

(Note, in general the i^{th} row of A is shifted left i-1 columns for i = 1, ..., n).

B. 1. The first column of B is left alone.

2. The second column of B is shifted up one row

3. The third column of B is shifted up two rows.

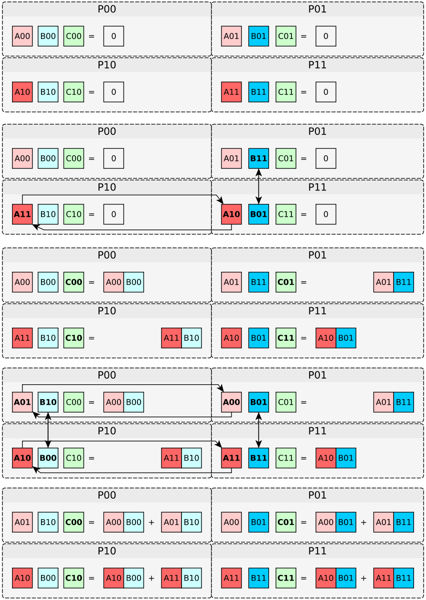
(Note, in general the j^{th} column of B is shifted up j-1 rows for j = 1, ..., n)

Once the registers have been shifted the multiplication pr

Demo

Cannon's Algorithm for Parallel Matrix Multiply: Demo for 2x2 block arrangment

$$\begin{bmatrix} C00 & C01 \\ C10 & C11 \end{bmatrix} = \begin{bmatrix} A00 & A01 \\ A10 & A11 \end{bmatrix} \times \begin{bmatrix} B00 & B01 \\ B10 & B11 \end{bmatrix}$$



1. Initial data layout.
Each Proc holds a block of A, B, and C respectively. Processors are arragned in a logical grid that reflects their initial data.

In subsequent steps, received / computed data is bolded.

2. Initial communication.
Row I: Rotate A blocks left by I procs
Col J: Rotate B blocks down by J procs

3. Multiply Step
Each Proc has a submatrix of A and B that allows partial computation of its output block of C. Stored blocks of A and B are multiplied and added onto contents of C.

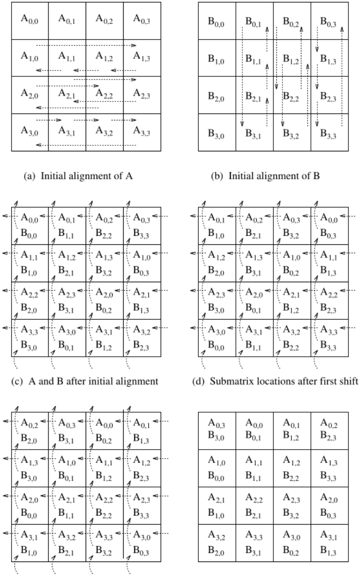
4. Subsequent communication.
Row I: Rotate A blocks left by 1.
Col J: Rotate B blocks down by 1.

5. Multiply Step (final).
Each Proc has a submatrix of A and B that allows partial computation of its output block of C

Algorithm terminates after sqrt(P) Communicate / Multiply steps.

Cannon's Algorithm Pseudocode

```
procedure Cannon(i, j, N){
  PE(i,j) has blocks A1=A(i,j) and B1=B(i,j)
  N is the Block Dimension : A is N*N blocks

  Allocate space A2, B2, Cij sized as A1

  doboth send A1 to   PE(i, j-i+N % N)
         recv A2 from PE(i, j+i+N % N)
  doboth send B1 to   PE(i-j+N % N, j)
         recv B2 from PE(i+j+N % N, j)

  for(k=1 to N){
    copy A2 into A1, B2 into B1
    Cij += A1 * B1

    doboth send A1 to   PE(i, j-1+N % N)
           recv A2 from PE(i, j+1+N % N)
    doboth send B1 to   PE(i-1+N % N, j)
           recv B2 from PE(i+1+N % N, j)
    // optionally skip last comm
  }

  Cij now contains output block of C(i,j)
}
```



(a) Initial alignment of A
(b) Initial alignment of B
(c) A and B after initial alignment
(d) Submatrix locations after first shift
(e) Submatrix locations after second shift
(f) Submatrix locations after third shift

Figure 8.3 The communication steps in Cannon's algorithm on 16 processes.

Exercise: Analysis of Cannon's Algorithm

Assumptions

- Matrices A and B are size $N \times N$ so N^2 elements
- P processors with block partitioning: initially $N^2/\sqrt{P}$ elements of A, B on each proc (assume P is a perfect square)
- Simplified communication cost for send/recv on a Ring:

$$t_{comm} = t_s + t_w M$$

with p #procs in ring, t_s comm startup time, t_w per word transfer rate, M message size.

Questions

1. What is communication cost of this Cannon's algorithm?
2. Is this an better/worse/same as the Naive algorithm?
3. What is the memory requirement for each proc?
4. Is this an better/worse/same as the Naive algorithm?

Answers: Analysis of Cannon's Algorithm

1. What is communication cost of this Cannon's algorithm?
   - In each step, each proc performs 2 send/recv ops
   - Each send/recv is a block of size $N^2/\sqrt{P}$
   - Total $\sqrt{P}$ steps : can skip last comm step

   $$t_{comm} = 2(\sqrt{P} - 1) \times (t_s + t_w(N^2/\sqrt{P}))$$

2. Is this an better/worse/same as the Naive algorithm?
   - Same communication cost as Naive algorithm
3. What is the memory requirement for each proc?
   - $O(N^2/\sqrt{P})$: 5 blocks as stated in pseudocode, 2 "workspaces" to allow send/recv of blocks
4. Is this an better/worse/same as the Naive algorithm?
   - Memory overhead is much better: constant number of blocks rather than the need to store entire rows/cols on single procs

Lessons from Cannon's Algorithm

- Illustrates "pipelining": blocks used to compute partial results then fed forward other processors
- Benefits greatly from a 2D Grid / Torus network which facilitates local communications that arise in the algorithm
- While not as ideal as row/col partitioning for A, B, realistic and relatively efficient
- Variants of central idea exist in some libraries such as Scalapack which has a parallel `xGEMM()` using many similar ideas
- Could really use some code support for
  - 2D Coordinates for processors rather than linear rank...
  - Sending/receiving in a ring...

MPI Tricks for Rings

Sendrecv in a Ring

`MPI_Sendrecv()` allows ring-link partnering

```
// sendrecv_ring.c
int left_part  = (myrank - 1 + npes) % npes;
int right_part = (myrank + 1 + npes) % npes;
MPI_Sendrecv(&mine,  1, MPI_INT, right_part, 1,
             &yours, 1, MPI_INT, left_part,  1,
             MPI_COMM_WORLD, MPI_STATUS_IGNORE);
```

Sendrecv with Replacement

`MPI_Sendrecv_replace()` allows send/recv in the same buffer

```
// sendrecv_ring.c
int mydata = 10*myrank;
MPI_Sendrecv_replace(&mydata, 1, MPI_INT,
                     right_part, 1, left_part,  1,
                     MPI_COMM_WORLD, MPI_STATUS_IGNORE);
```

In Cannon's Alg, no longer need `A1 / A2`: can send/receive block of `A` with a single buffer.

MPI Tricks for Grids: `MPI_Cart_create()`

MPI has special support for Grid/Torus network configs; allows creation of a `MPI_Comm` that maps processors to a N-D grid

- 2D Torus for Cannon's Alg

```
// cartesian_comm.c
int dim_len = 2;  // Set up the Cartesian topology
int dims[2] = {sqrt(npes), sqrt(npes)}; // # rows/cols
int periods[2] = {1, 1};      // wrap-around rows/cols

// Create the Cartesian topology, with rank reordering
MPI_Comm comm_2d;
MPI_Cart_create(MPI_COMM_WORLD,         // original comm
                dim_len, dims, periods, // cartesian comm props
                1,           // re-order linear rank if beneficial
                &comm_2d);   // new communicator with 2D coords

// Get the rank and coordinates with respect to the new topology
int my2drank = -1;           // may be differ from world rank
MPI_Comm_rank(comm_2d, &my2drank);

int mycoords[2] = {-1, -1};   // (i,j) coords
MPI_Cart_coords(comm_2d, my2drank, 2, mycoords);

printf("Proc %2d (%s): my2drank %3d mycoords (%3d, %3d)\n",
        myrank,processor_name,
        my2drank,mycoords[0],mycoords[1]);
```

MPI Tricks for Shifting

Shifts are eased by the `MPI_Cart_shift()` function

- Calculates linear rank of source/dest procs for shift operations in a Cartesian grid of procs.
- Data exchange via `MPI_Sendrecv()` is then direct

```
// cartesian_comm.c
int mydata = (100*mycoords[0])+mycoords[1];
int rowsend=-1, rowrecv=-1;

MPI_Cart_shift(comm_2d, 0, rowshift, &rowrecv, &rowsend);

MPI_Sendrecv_replace(&mydata, 1, MPI_INT,
                     rowsend, 1, rowrecv,  1,
                     MPI_COMM_WORLD, MPI_STATUS_IGNORE);
```

Cannon's Algorithm in MPI

- Grama Program 6.2 is Cannon's Matrix Multiply algorithm implemented via MPI
- Uses the tricks mentioned on the past 2 slides to ease implementation burden
- See `cannon_grama.c` for a source code version of it

Note: I haven't tested this code but everything from textbooks always works out the box, right?

Next Time: Solving Linear Systems via LU Decomposition