Blocked In-Place Rectangular Transpose

Fred Gustavson
Assisted by Lars Karlsson

F13-part2: Design & Analysis of Algorithms for HPC Systems; Umeå University, Sweden

May 20, 2008

Main Idea

- Combine Cache Blocking with Point In-Place Transpose on a very tiny matrix
  - use of RB Format is the key idea
  - CM -> RB -> RB^T -> CM
  - SB is special case of RB
- Block In-Place Transpose is Very Fast relative to Point In-Place Transpose
- CM <-> RB uses fast vector In-Place Alg.

Summary or Overview

- A is M by N.
- M=m*MB & N=n*NB
- CM -> RB by vector IP transpose
- RB <-> RB^T by block IP transpose
  - use point IP transpose on m by n A1= SB A
- RB^T -> CM by vector IP transpose

Vector In-Place Xpose or CM<->SB

- Let A be M by NB with M=m*MB
- View A as m by NB A1 with each a1(i,j) being a column vector of size NB
- Apply point IP transpose to A1 to get A2
- A2 is m order NB SB's concatenated
- Apply above subroutine n=N/NB times

Where does the Speed Come From

- Data moved in blocks and vectors gives a 10 to 100 times performance gain
  - uses stride one processing; every line gets fully used when it enters L1 and streaming by algorithmic / automatic pre-fetching works
- SMP parallelism is easy to implement
  - disjoint cycle structure
  - long cycles can be broken into pieces

Other Matrix Layouts

- Can block transform (in-place) any permutation that can be described by a compact functional description
  - includes all common matrix data layouts
    - standard CM / RM rectangular arrays
    - standard CM / RM triangular arrays
    - standard packed format

An Example of CM to RB

- A is M = 500 by NB = 4
  - M = m*MB with MB = 100 and m = 5
- A1 is m = 5 by NB = 4; each element of A1 is a vector of length MB = 100
- Both A and A1 are identical in storage and occupy M*NB = 2000 contiguous locations

Picture of the Previous slide

	0	500	1000	1500
	0	1	2	3
0	00	01	02	03
100	10	11	12	13
200	20	21	22	23
300	30	31	32	33
400	40	41	42	43

	0	400	800	1200	1600
	0	1	2	3	4
0	00	10	20	30	40
100	01	11	21	31	41
200	02	12	22	32	42
300	03	13	23	33	43

Previous slide shows CM to RB is vector in place transpose

- The left matrix A1 is a m = 5 by NB = 4 matrix whose elements are vectors of length MB = 100. This matrix is in standard CM format.
- The right matrix is the NB = 4 by m = 5 vector transpose of A1. It is in RB format consisting of m = 5 RB’s of size MB = 100 by NB = 4.

Details of vector in place transpose

- 0 and m*n -1 = 10 are singleton cycles
- 19 is prime and # d = 2; 1 & 19
- q = m*n – 1 = 19 is the mod value
- For problem 19, phi = 18 & cl = 9; leaders are 1, 2
- For problem 1, phi = 1 & cl = 1 at 19
- Further details follow

More Details continued

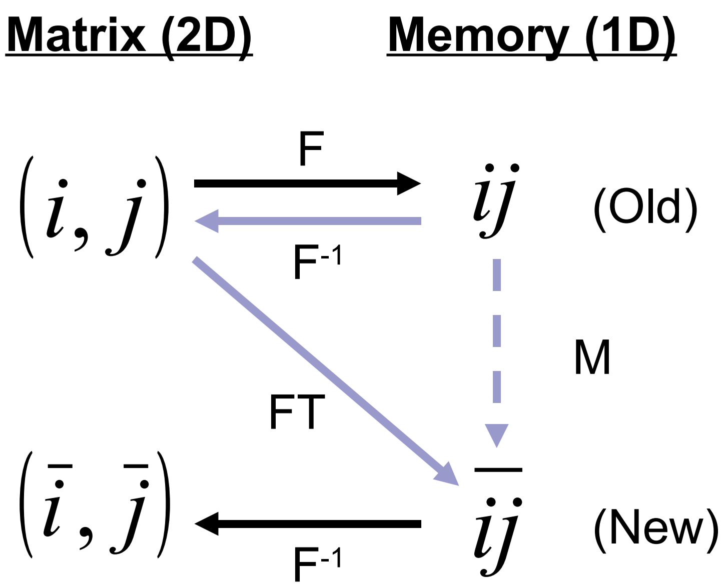
- cycle one = 1, 4, 16, 7, 9, 17, 11, 6, 5, 1
- cycle two =2, 8, 13, 14, 18, 15, 3, 12, 10, 2
- cycle three = 19
- cycle four = 0
- These four cycles cover all of A1’s twenty vectors of length MB = 100
- These four cycles cover all 2000 elements of A and transform A from CM format to RB format

A1 layout of A; see slide # 8

	0	5	10	15
	0	1	2	3
0	00	01	02	03
1	10	11	12	13
2	20	21	22	23
3	30	31	32	33
4	40	41	42	43

	0	4	8	12	16
	0	1	2	3	4
0	00	10	20	30	40
1	01	11	21	31	41
2	02	12	22	32	42
3	03	13	23	33	43

In Place Transpose Mapping



$$F(i,j) \mapsto i + m \cdot j$$

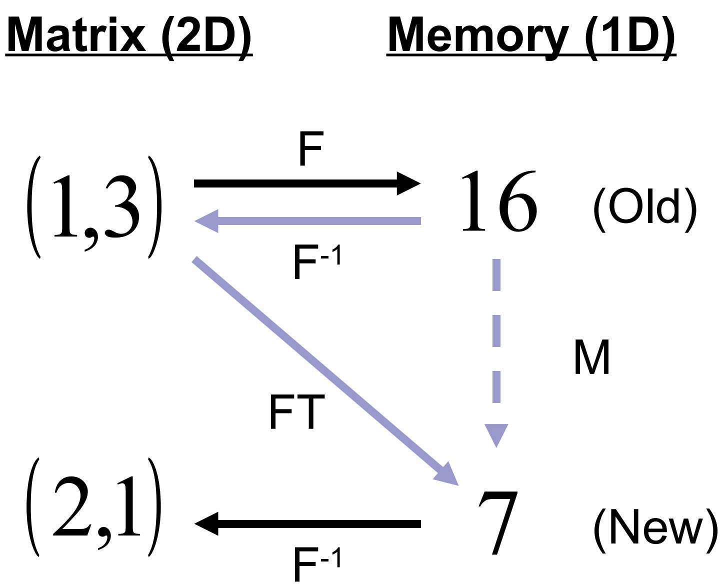
$$F^{-1}(ij) \mapsto \left(\operatorname{mod}(ij, m), \left\lfloor \frac{ij}{m} \right\rfloor \right)$$

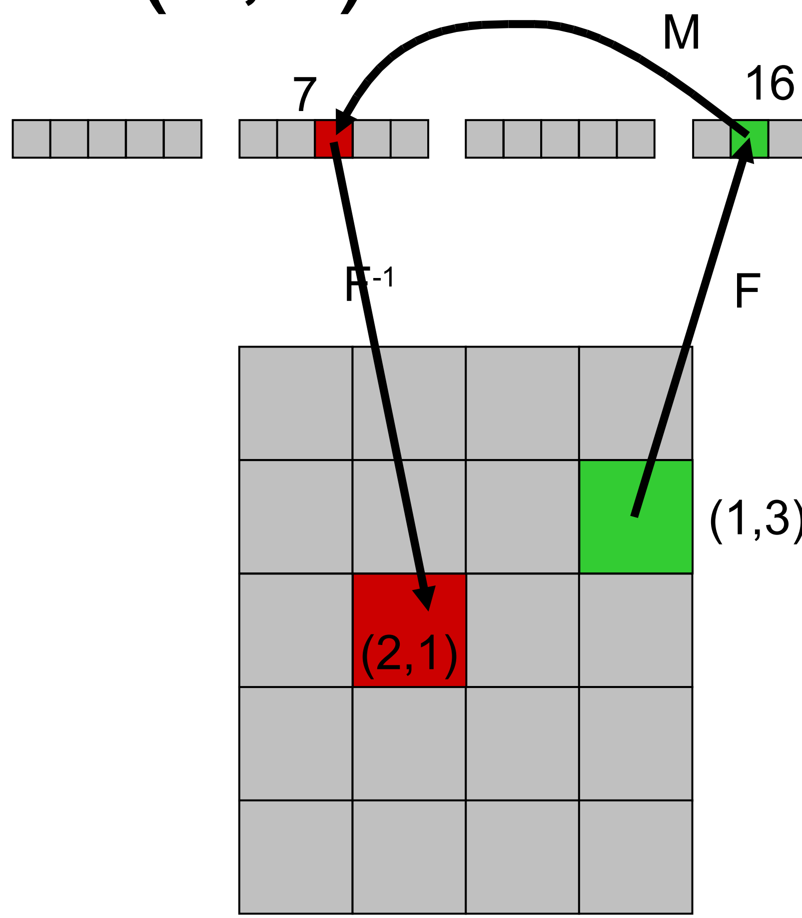
$$FT(i,j) \mapsto j + n \cdot i$$

$$M(ij) \mapsto \operatorname{mod}(n \cdot ij, q)$$

$$M = FT \circ F^{-1}$$

Mapping (1,3) to (2,1)





Picture of the 4 cycles of slide # 11

0	8	8	5
0	7	6	2
0	3	7	5
6	1	2	4
1	4	3	0

A is 500 by 700 in CM order

- CM A has LDA = 500
- A has 7 column swaths: 500 by 100 each
- A1 is 5 by 100 matrix of vectors
- In-place transpose with q = 499
- repeat above 6 more times
- A is now in SB format of size 5 by 7

Details of CM to SB Vector

- 0 and m*n -1 = 499 are singleton cycles
- 499 is prime and # d = 2; 1 & 499
- q = m*n – 1 is the mod value
- for problem 499, phi = 498 & cl = 249; leaders are 1, 2
- for problem 1, phi = 1 & cl = 1 at 499

Details of SB to SB^T

- q = 5*7 – 1 = 34 = 2*17
- q = sum over divisors of phi
  - #d = 4; 34, 17, 2, 1; phi's = 16, 16, 1, 1
- #d problems gives cycles of length 16, 16, 1, 1 starting at 1, 2, 17, 34

Cycles for SB to SB^T

- Map (ij) = mod (ij*n, q) ; m = 5, n = 7, q = m*n – 1 = 34
- cycle one : 1, 7, 15, 3, 21, 11, 9, 29, 33, 27, 19, 31, 13, 23, 25, 5, 1
- cycle two : 2, 14, 30, 6, 8, 22, 18, 24, 32, 20, 4, 28, 26, 12, 16, 10, 2
- cycles at 17, 34, and 0 are singletons

Picture of Map on slide # 19

0	15	15	2	9	14	2
0	3	5	14	4	12	11
0	1	13	0	5	9	8
3	4	12	6	13	11	8
10	6	1	10	7	7	0

Details of SB^T to CM

- m = 100, n = 7, q = 699 = 3*233
- # d = 4; 699, 233, 3, 1; phi’s 464, 232, 2, 1
- cl’s are 166, 166, 1, 1
- leaders are 1, 2, 5, 10; 3, 9; 233, 466; 699

The 500 by 700 A as a point matrix

- q = m*n – 1 = 349,999 = 13**2*19*109
- # d's = 3*2*2 = 12:
- sum of phi(d) = q
- twelve phi's are 303264, 23328, 16848, 2808, 1944, 1296, 216, 156, 108, 18, 12, 1
- twelve cl's are 468, 36, 156, 468, 18, 12, 36, 156, 6, 9, 12, 1
- ratio's give # of leaders: 648, 648, 108, 6,108, 108, 6, 1, 18, 2, 1, 1: sum = 1655

500 by 700 A as point matrix

- hand-out has cycle of length 12 at ij=247
  - ij is (247,0) element of A; next element in cycle is mod(247*700,q); 247 | q so we get cycle is i <- mod(700*i,1417) :
  - 247*(1, 700, 1135, 980, 172, 1372, 1091, 1354, 1244, 762, 608, 500, 1) mod (q)