

Sec 2. Memory Architecture cont'd

Memory Organizations and Cache



1

UNIVERSITY of
Rhode Island

Example Program Optimizing Cache Performance

```

1. j := 1;
2. for jj = 1 to N2/B2 do begin /* jj-loop */
3.   for kk = 1 to N/B1 do /* kk-loop */
4.     for i = 1 to N1 do /* i-loop */
5.       for k = B1(kk-1) + 1 to kkB1 do /* k-loop */
6.         begin
7.           put X[i,k] in a register;
8.           for ii = j to j + B2 - 1 do /* ii-loop */
9.             Z[i,ii] := Z[i,ii] + X[i,k] * Y[k,ii];
10.          end;
11.        j := j + B2;
12.      end jj;

```



2

UNIVERSITY of
Rhode Island

Example Program Optimizing Cache Performance

- A matrix multiplication algorithm
- The total size of the matrix may be larger than cache size
- To better utilize memory, blocking is generally used

$$\begin{pmatrix} X[N1 \times N] \end{pmatrix} \times \begin{pmatrix} Y[N \times N2] \end{pmatrix} = \begin{pmatrix} Z[N1 \times N2] \end{pmatrix}$$

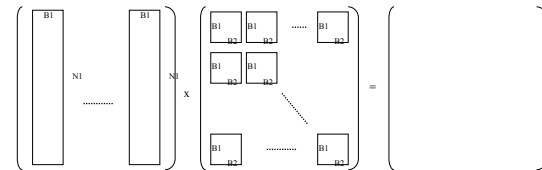


3

UNIVERSITY of
Rhode Island

Blocking the Matrix

- Partition Matrix Y into B1 by B2 blocks
- Once a block of Y is loaded in the cache, reuse it for maximum times

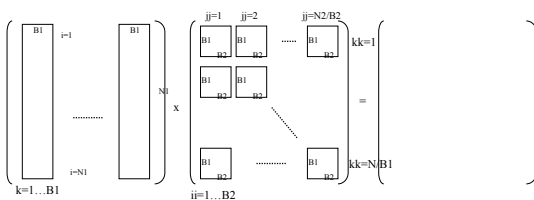


4

UNIVERSITY of
Rhode Island

Blocking the Matrix

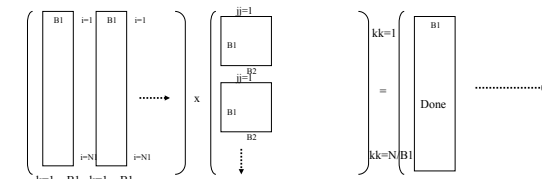
- Partition Matrix Y into B1 by B2 blocks
- Once a block of Y is loaded in the cache, reuse it for maximum times



5

UNIVERSITY of
Rhode Island

Finish all computations with one B1xB2 block

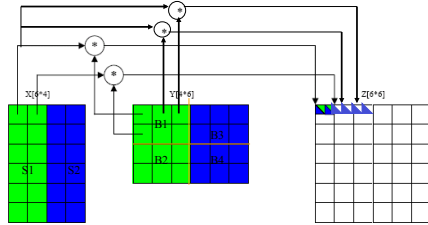


6

UNIVERSITY of
Rhode Island

An Example of $[6 \times 4] \times [4 \times 6]$

- Y is partitioned into 2 blocks of 4 by 3
- We do one column of X at a time
- Accumulate partial sum



7

UNIVERSITY of
Rhode Island

Partial Results of The first block

- Partial results are accumulated in the first 6 by 3 block of Z

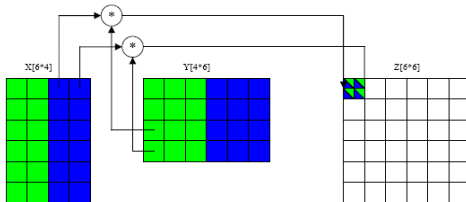


8

UNIVERSITY of
Rhode Island

More accumulations

- We continue on the remaining columns of X and remaining rows of Y

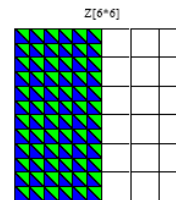


9

UNIVERSITY of
Rhode Island

Complete Accumulation of the First Block of Z

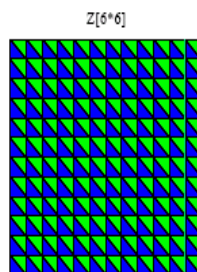
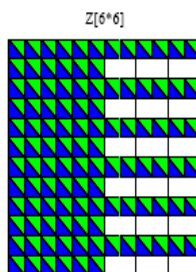
- We finish accumulating the result in the first block of Z



10

UNIVERSITY of
Rhode Island

The Same Process Continues on the Other Half



11

UNIVERSITY of
Rhode Island