CPEG 422/622
EMBEDDED SYSTEMS DESIGN

Chengmo Yang

chengmo@udel.edu

Evans 201C

LECTURE 15
PROJECT 4

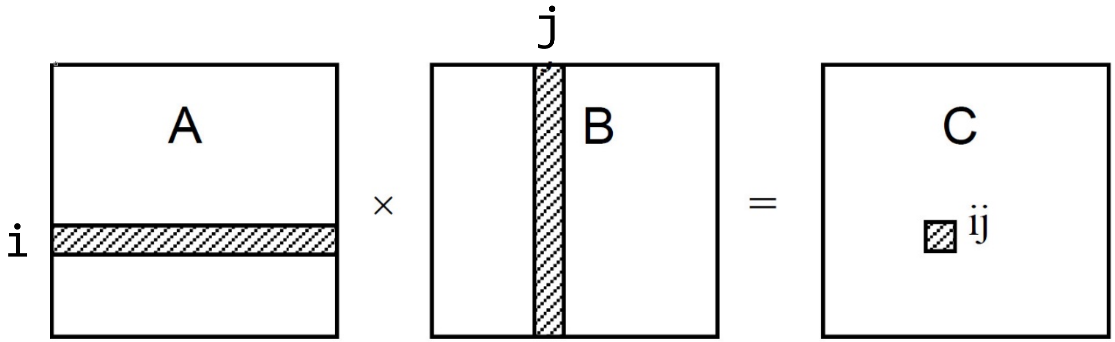
OUTLINE

- Matrix multiplication introduction
- Project 4

MATRIX MULTIPLICATION

Sequential matrix multiplication:

```
for $i$ = 0 to $m$ – 1 do
   for $j$ = 0 to $m$ – 1 do
       $c_{ij}$ := 0
     for $k$ = 0 to $m$ – 1 do
       $c_{ij}$ :+= $a_{ik}b_{kj}$
     endfor
    endfor
 endfor
```



MATRIX MULTIPLICATION

- Why suitable for the FPGA?
  - Number of array elements transferred: $3n^2$
  - Number of multiplications performed: n^3
  - Number of additions performed: n^3
  - Make sense for large number of n
- Better algorithms that improve slightly
  - Multiplication by block
  - Use distributed shared-memory multiprocessor with caches

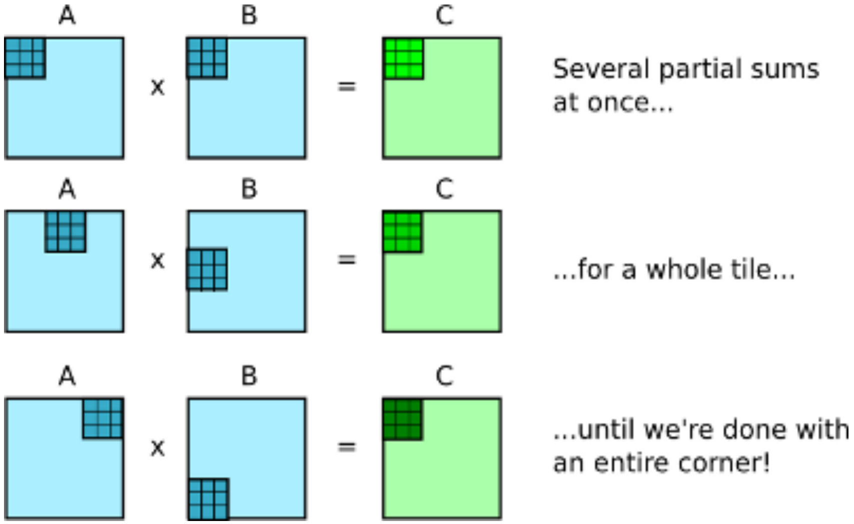
MORE EFFICIENT MATRIX MULTIPLICATION

Consider A,B,C to be N by N matrices of b by b subblocks where b is called the block size.

Example:

matrix is divided into 3 by 3 subblocks. The three partial products can be computed in parallel.

```
for i = 1 to N for j = 1 to N
   {read block C(i,j) into fast memory}
 for k = 1 to N
   {read block A(i,k) into fast memory}
   {read block B(k,j) into fast memory}
    C(i,j) = C(i,j) + A(i,k) * B(k,j)
  {do a matrix multiply on blocks}
  {write block C(i,j) back to slow memory}.
```



Several partial sums at once...

...for a whole tile...

...until we're done with an entire corner!

PROJECT 4

- Task 1: implement basic sequential matrix multiplication in software.
- Task 2: implement a fixed 3 by 3 size matrix multiplication in hardware.
- Task 3: implement a scalable matrix multiplication in hardware, and find the maximum matrix size that can be held on FPGA.
- Task4: compare running times of software and hardware implementations for different matrix sizes.
- Task 5: write a comprehensive experiment report.

PROJECT 4

Minimal Hardware and Toolkit:

- Vivado 2017.4
- Zybo
- Xillinx SDK
- Serial terminal

PROJECT 4

- Questions:

1.Please describe how to make your VHDL code being scalable for different matrix sizes.

2. Please numerate possible hardware bottleneck types in term of resource category.

For each possible bottleneck you find, what is the maximum matrix size can be held on FPGA?

In your design, what is the real bottleneck, and the maximum matrix size can be held in your design?

(Hint: resource types such as LUT, RAM…)

PROJECT 4

Submission:

- Demo and report are required.

Due dates:

- Code: May 18th noon
- In-class demo: May 18th
- Report: May 20th midnight