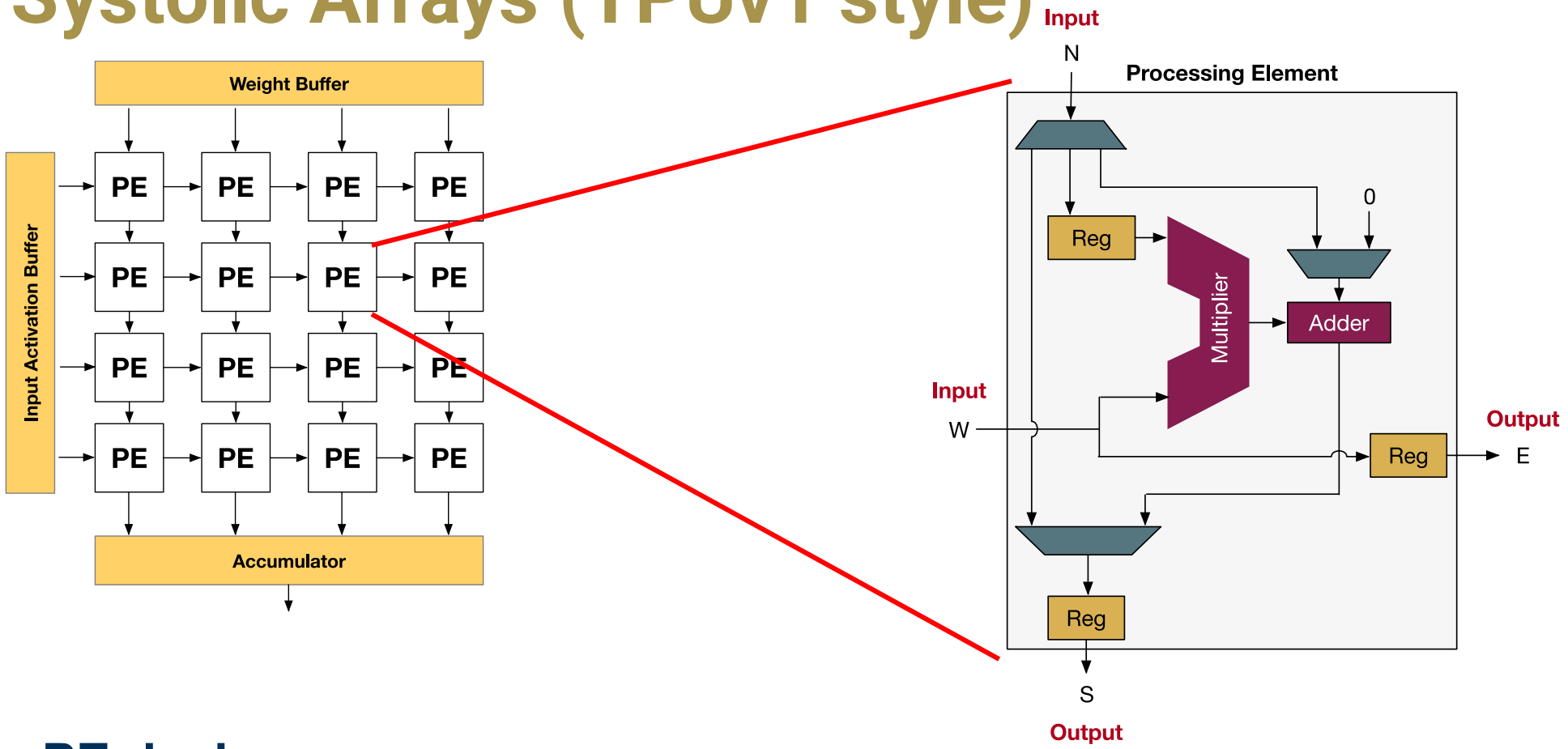
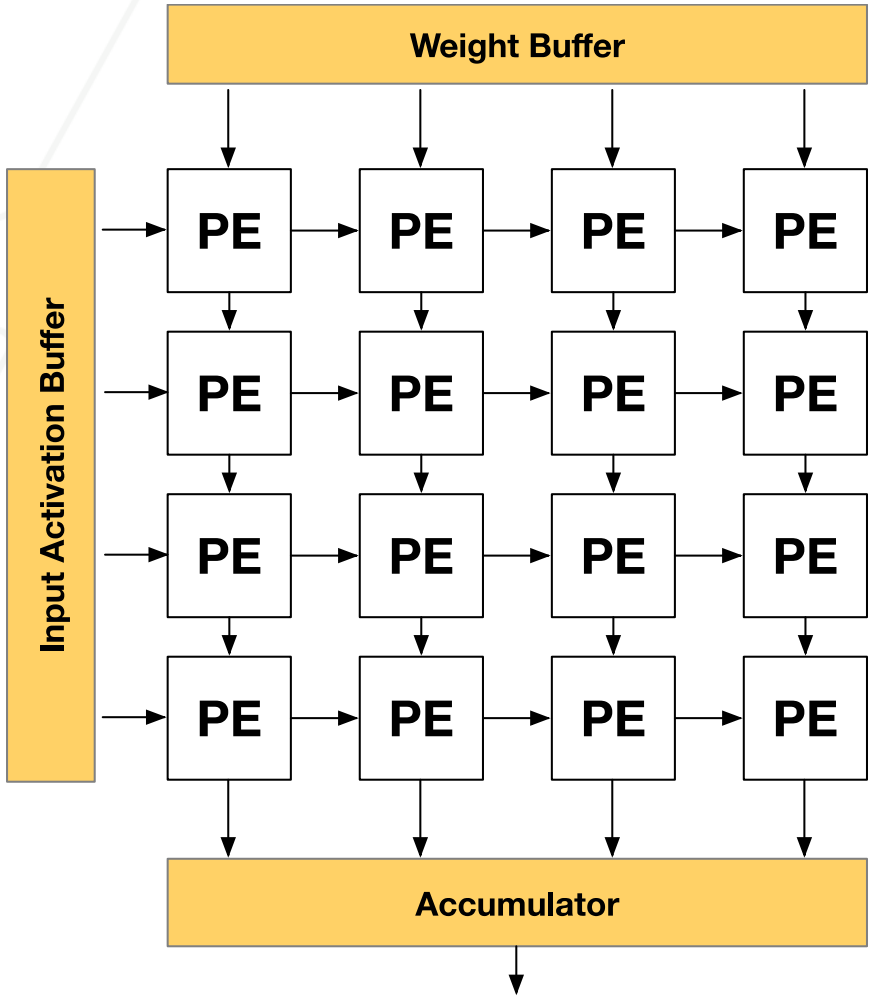
PEs in Systolic Arrays (TPUv1 style)

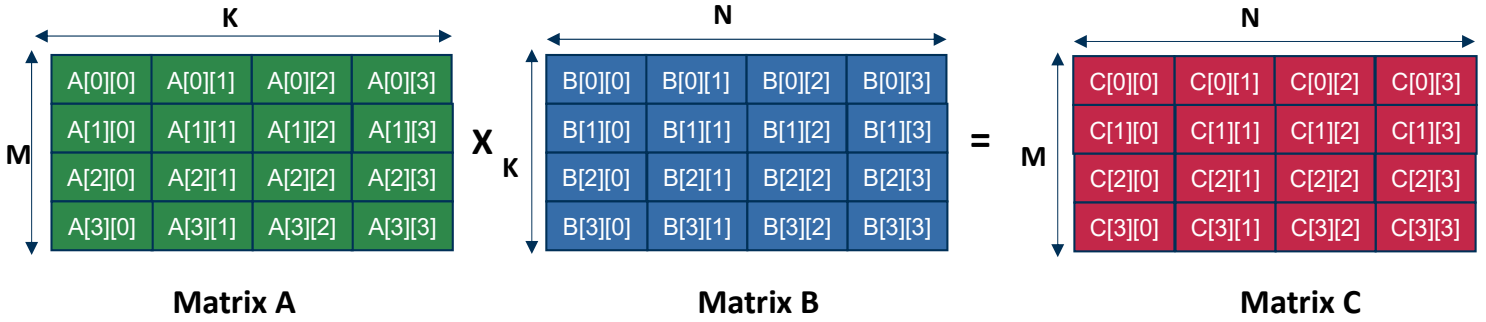


- **Simple PE design**
  - No local scratchpad; several one-slot registers
  - Enables to increase the number of PEs (small area for each PE -> more PEs within the same area)

Georgia Tech

Matrix Multiplication on a Systolic Array






```
for m in range(4):
  for n in range(4):
    for k in range(4):
      C[m][n] += A[m][k] * B[k][n]
```

Assume following "mapping"

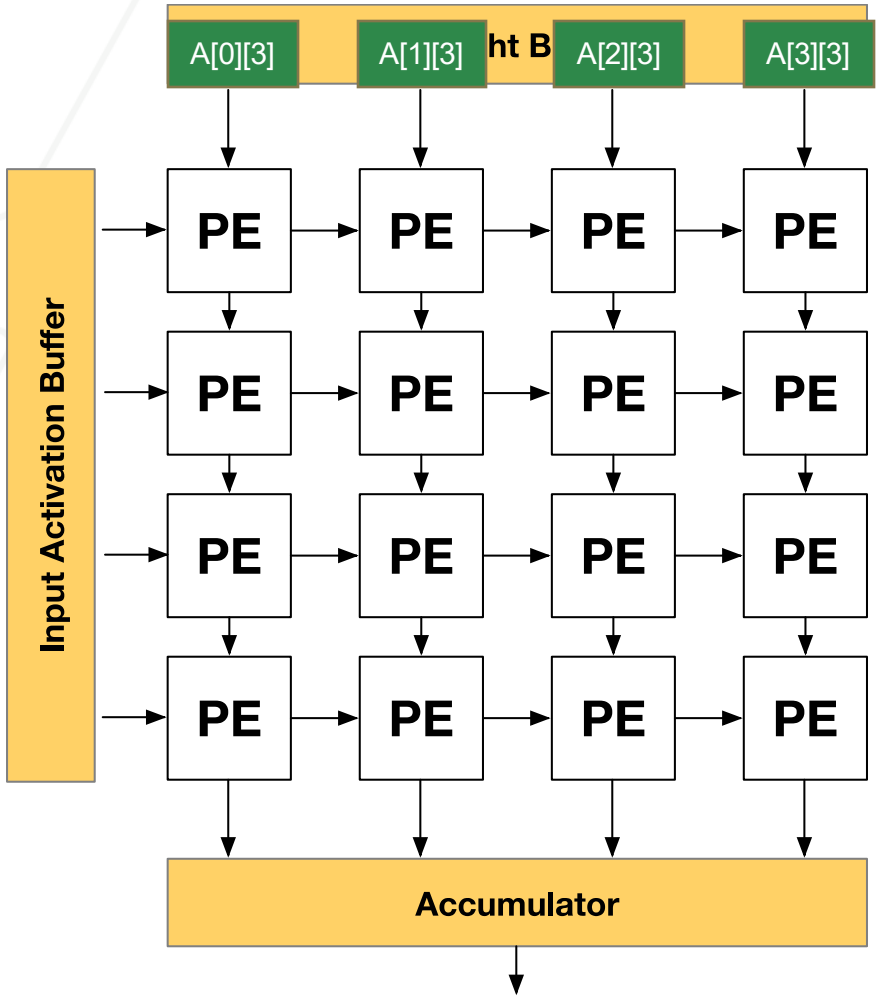
- Column: Reduced (*aka* contracted) Dimension (K)
- Row: Spatial Dimension (either M or N)

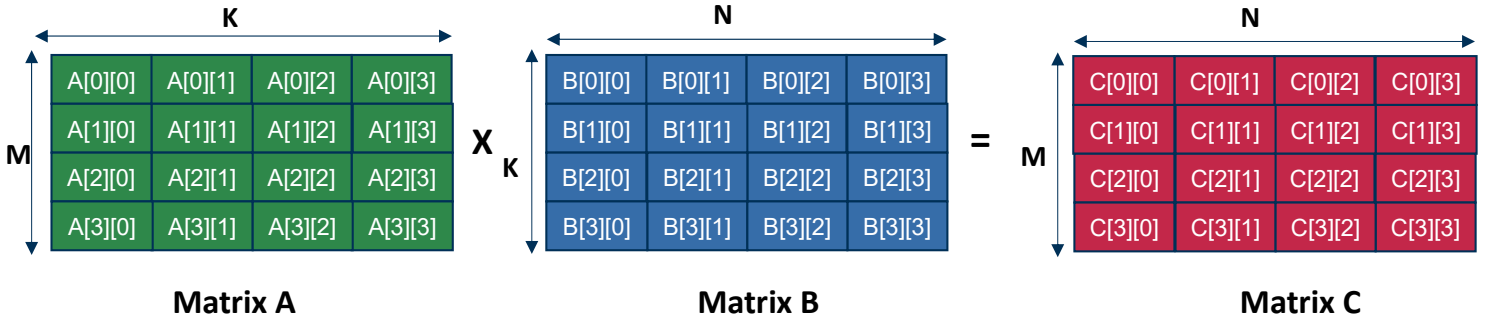
Note: alternate mapping styles possible

2

Matrix Multiplication on a Systolic Array

Phase 1:
Load Weight Tensor






```
for m in range(4):
  for n in range(4):
    for k in range(4):
      C[m][n] += A[m][k] * B[k][n]
```

Assume following "mapping"

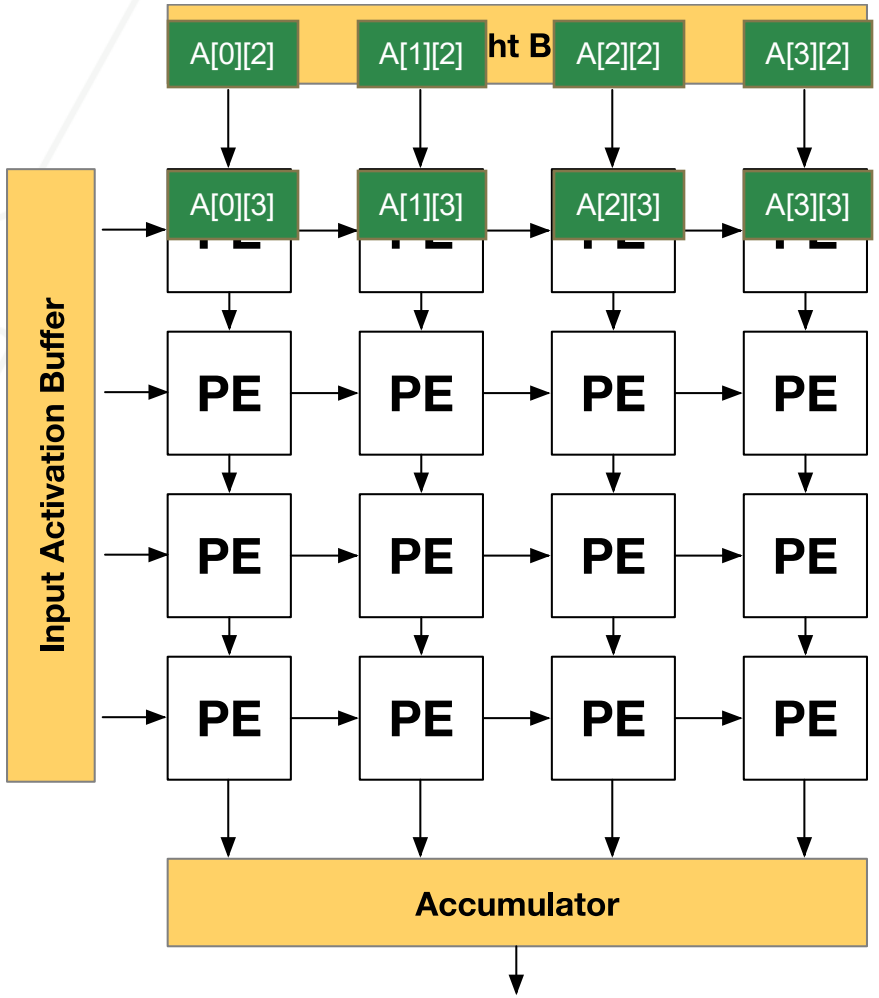
- Column: Reduced (*aka* contracted) Dimension (K)
- Row: Spatial Dimension (either M or N)

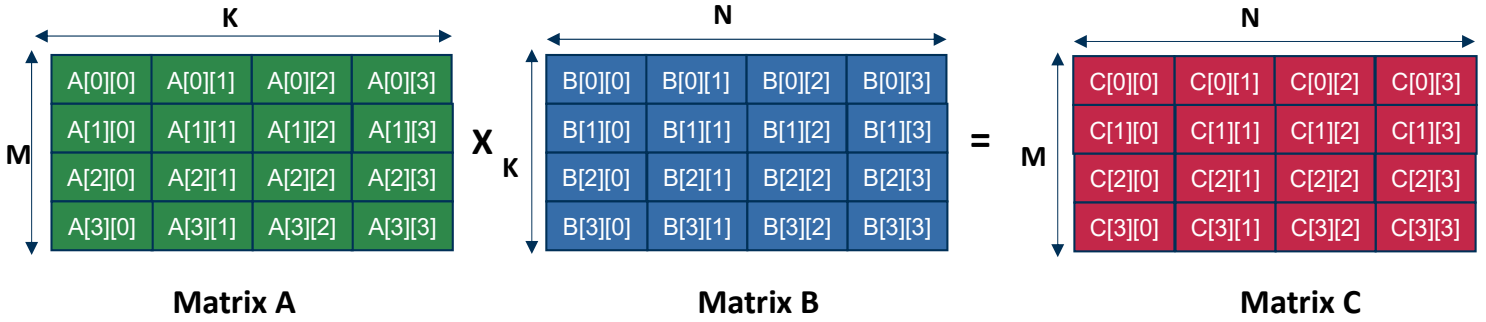
Note: alternate mapping styles possible

Georgia Tech

Matrix Multiplication on a Systolic Array

Phase 1:
Load Weight Tensor






```
for m in range(4):
  for n in range(4):
    for k in range(4):
      C[m][n] += A[m][k] * B[k][n]
```

Assume following "mapping"

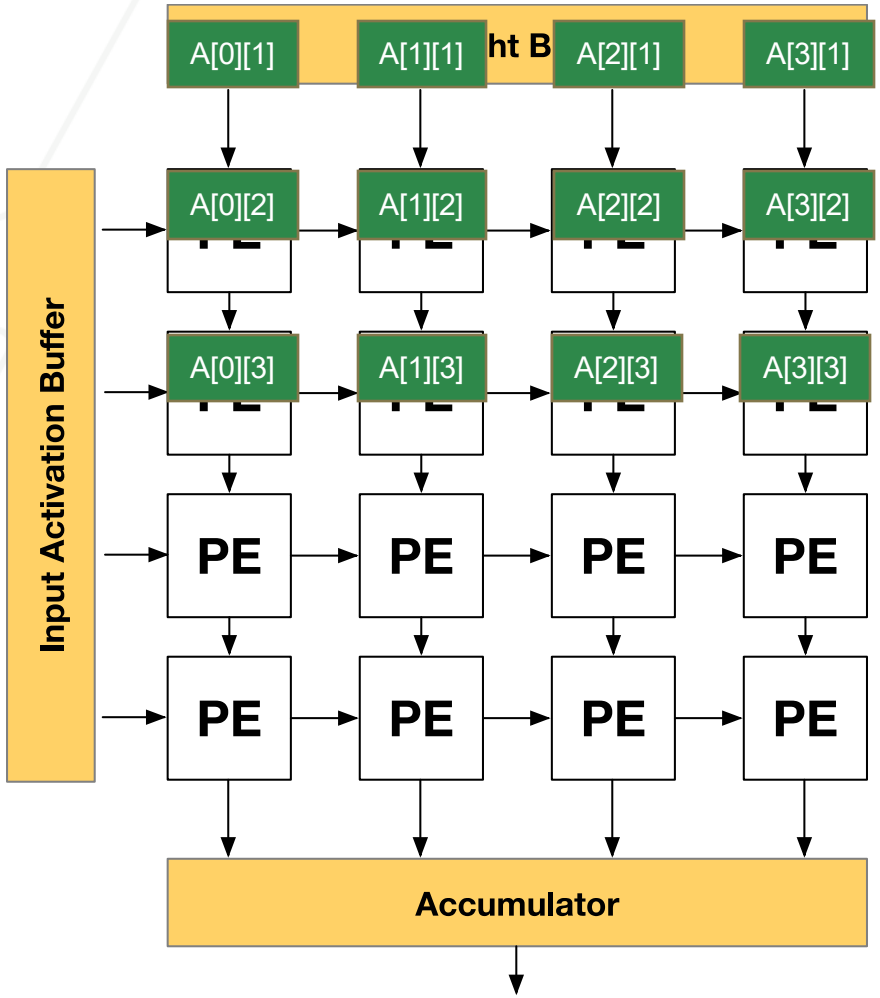
- Column: Reduced (*aka* contracted) Dimension (K)
- Row: Spatial Dimension (either M or N)

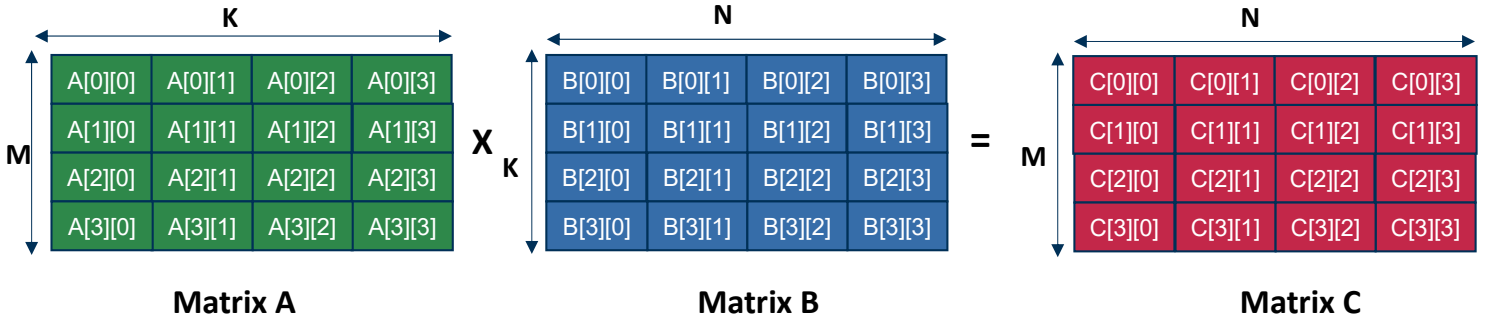
Note: alternate mapping styles possible

Georgia Tech

Matrix Multiplication on a Systolic Array

Phase 1:
Load Weight Tensor






```
for m in range(4):
  for n in range(4):
    for k in range(4):
      C[m][n] += A[m][k] * B[k][n]
```

Assume following "mapping"

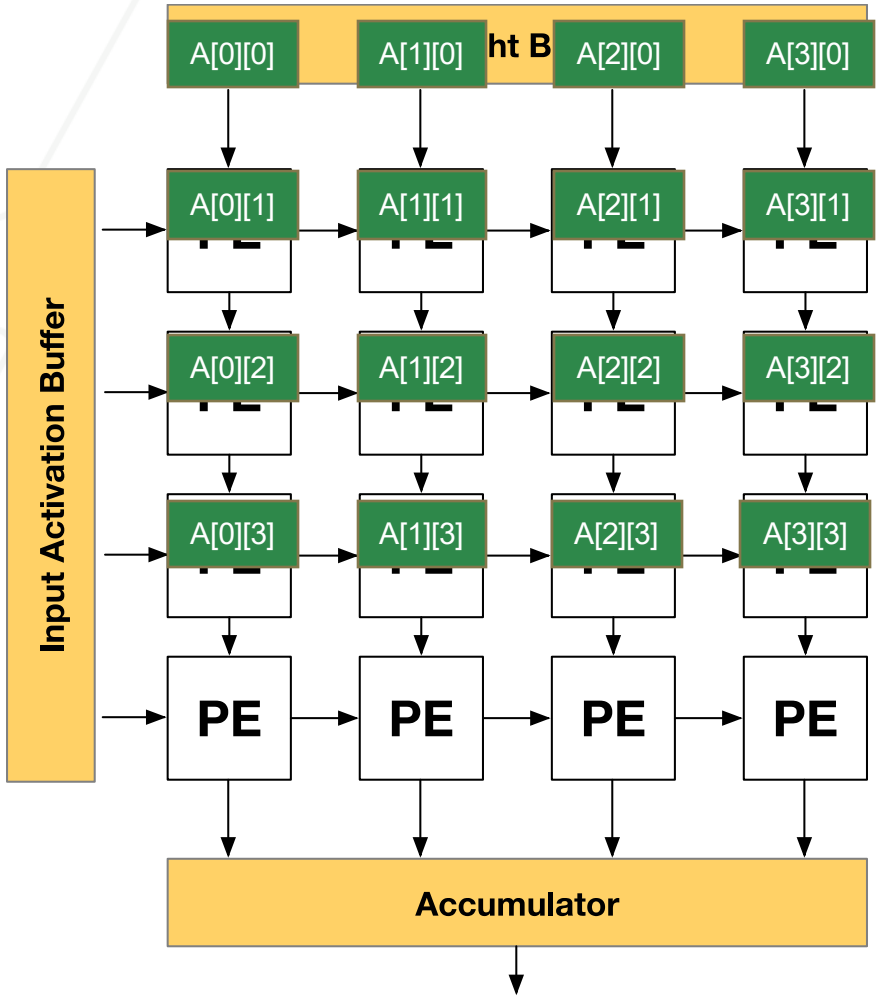
- Column: Reduced (*aka* contracted) Dimension (K)
- Row: Spatial Dimension (either M or N)

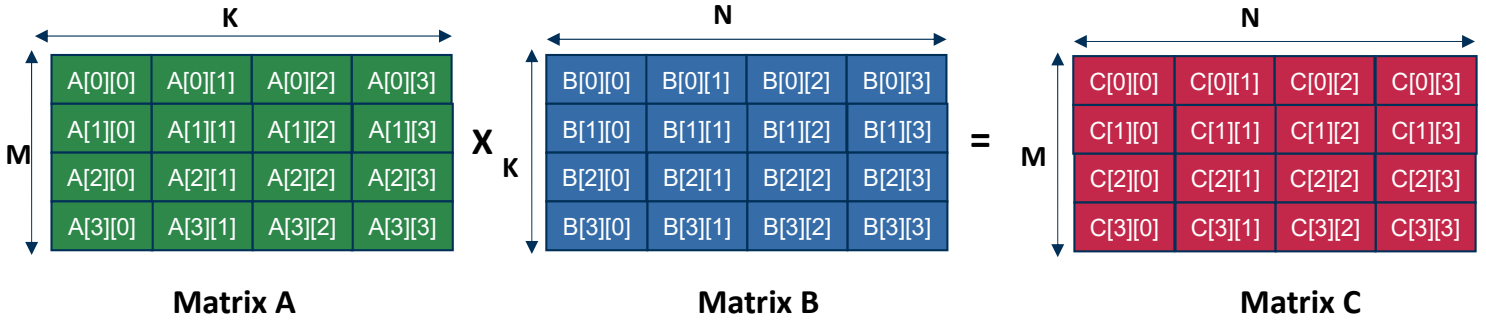
Note: alternate mapping styles possible

5

Matrix Multiplication on a Systolic Array

Phase 1:
Load Weight Tensor






```
for m in range(4):
  for n in range(4):
    for k in range(4):
      C[m][n] += A[m][k] * B[k][n]
```

Assume following "mapping"

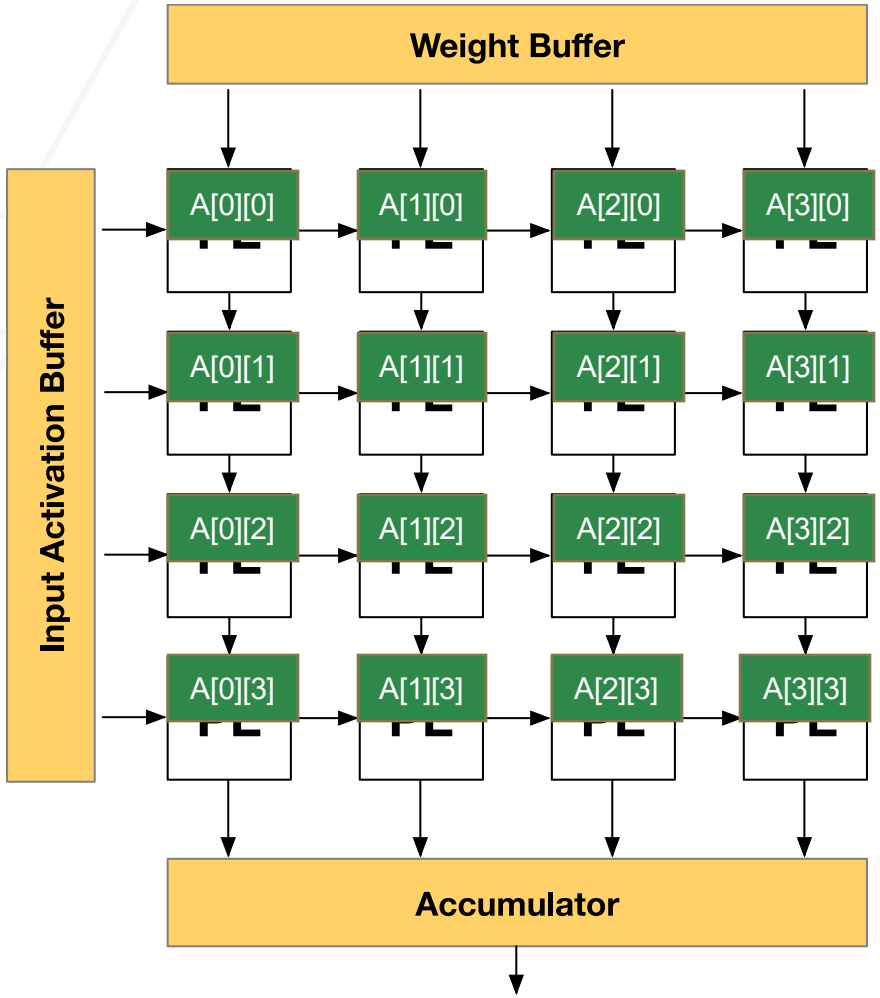
- Column: Reduced (*aka* contracted) Dimension (K)
- Row: Spatial Dimension (either M or N)

Note: alternate mapping styles possible

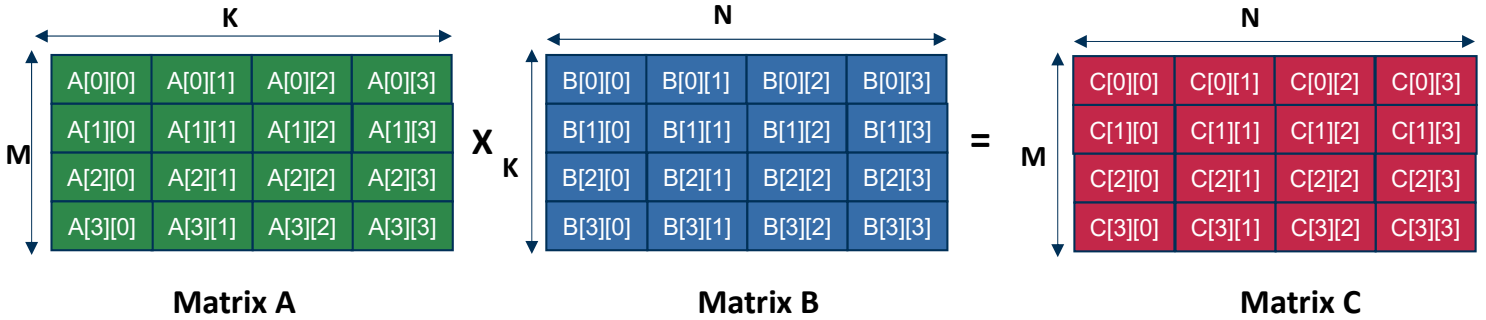
6

Matrix Multiplication on a Systolic Array

Phase 1:
Load Weight Tensor



Matrix A will now stay resident (i.e., "stationary") in the PEs




```
for m in range(4):
  for n in range(4):
    for k in range(4):
      C[m][n] += A[m][k] * B[k][n]
```

Assume following "mapping"

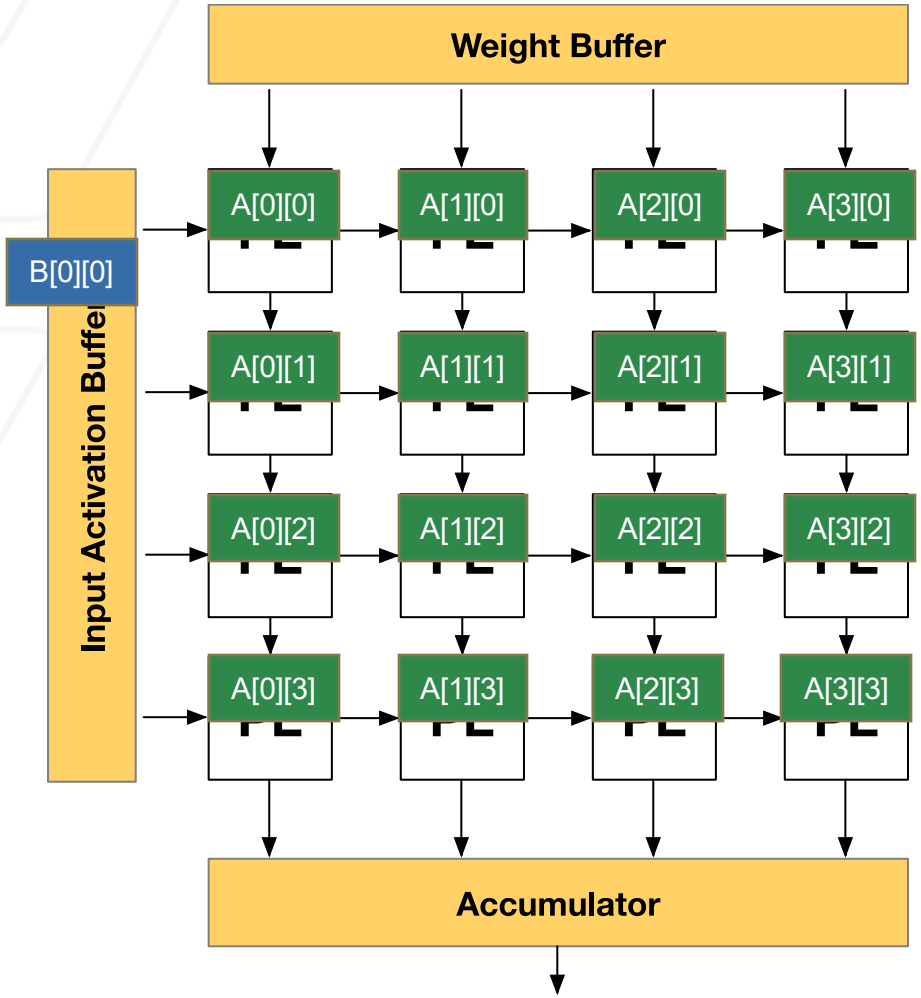
- Column: Reduced (*aka* contracted) Dimension (K)
- Row: Spatial Dimension (either M or N)

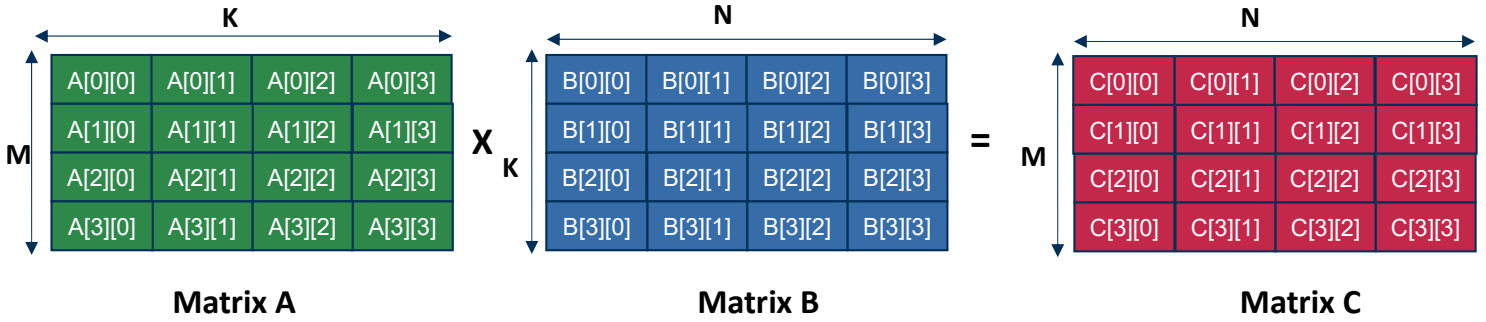
Note: alternate mapping styles possible

7

Matrix Multiplication on a Systolic Array

Phase 2:
Stream Activation Tensor






```
for m in range(4):
  for n in range(4):
    for k in range(4):
      C[m][n] += A[m][k] * B[k][n]
```

Assume following "mapping"

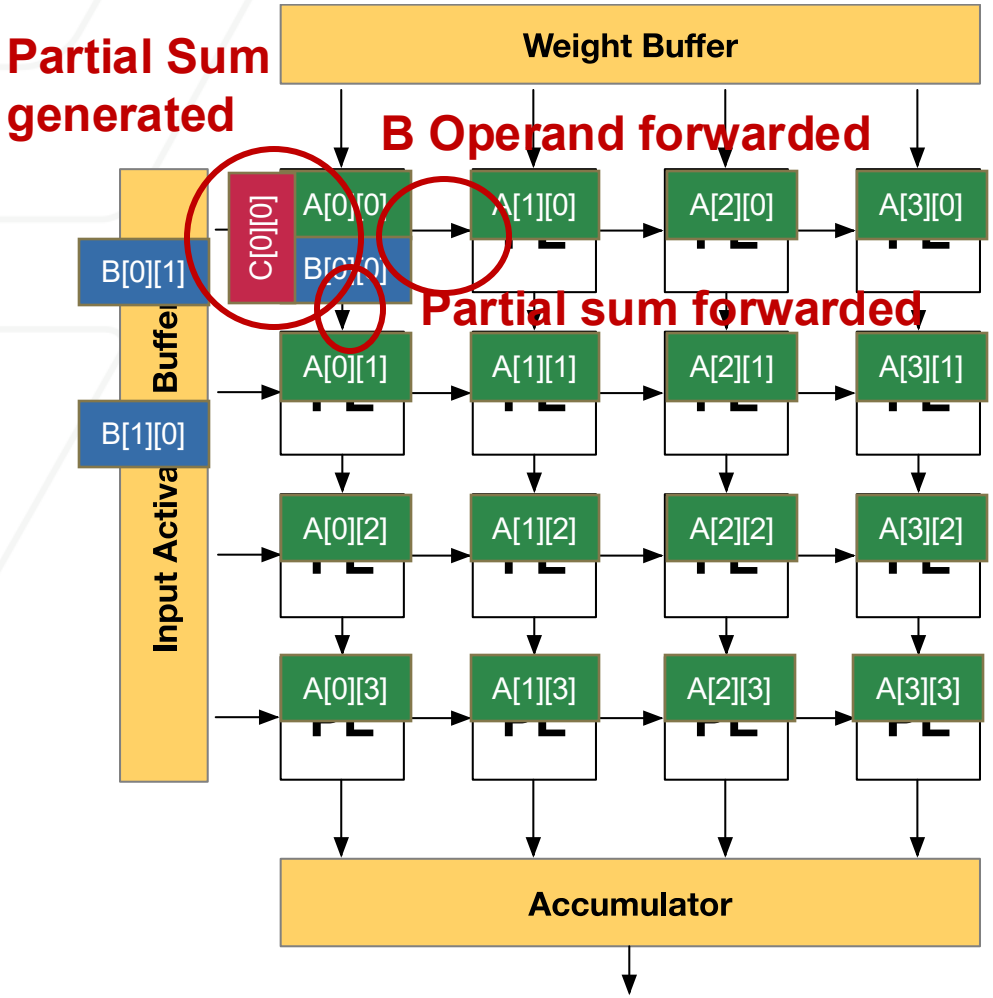
- Column: Reduced (*aka* contracted) Dimension (K)
- Row: Spatial Dimension (either M or N)

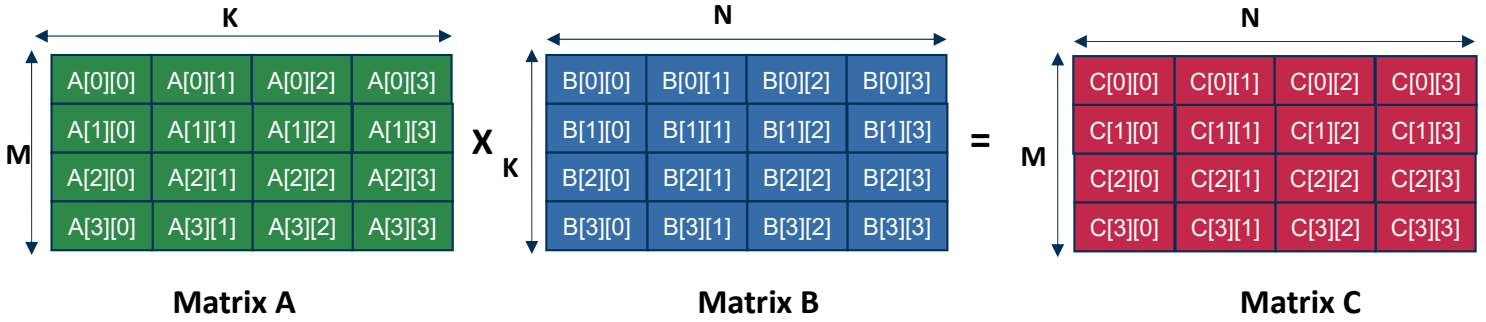
Note: alternate mapping styles possible

Georgia Tech

Matrix Multiplication on a Systolic Array

Phase 2:
Stream Activation Tensor






```
for m in range(4):
  for n in range(4):
    for k in range(4):
      C[m][n] += A[m][k] * B[k][n]
```

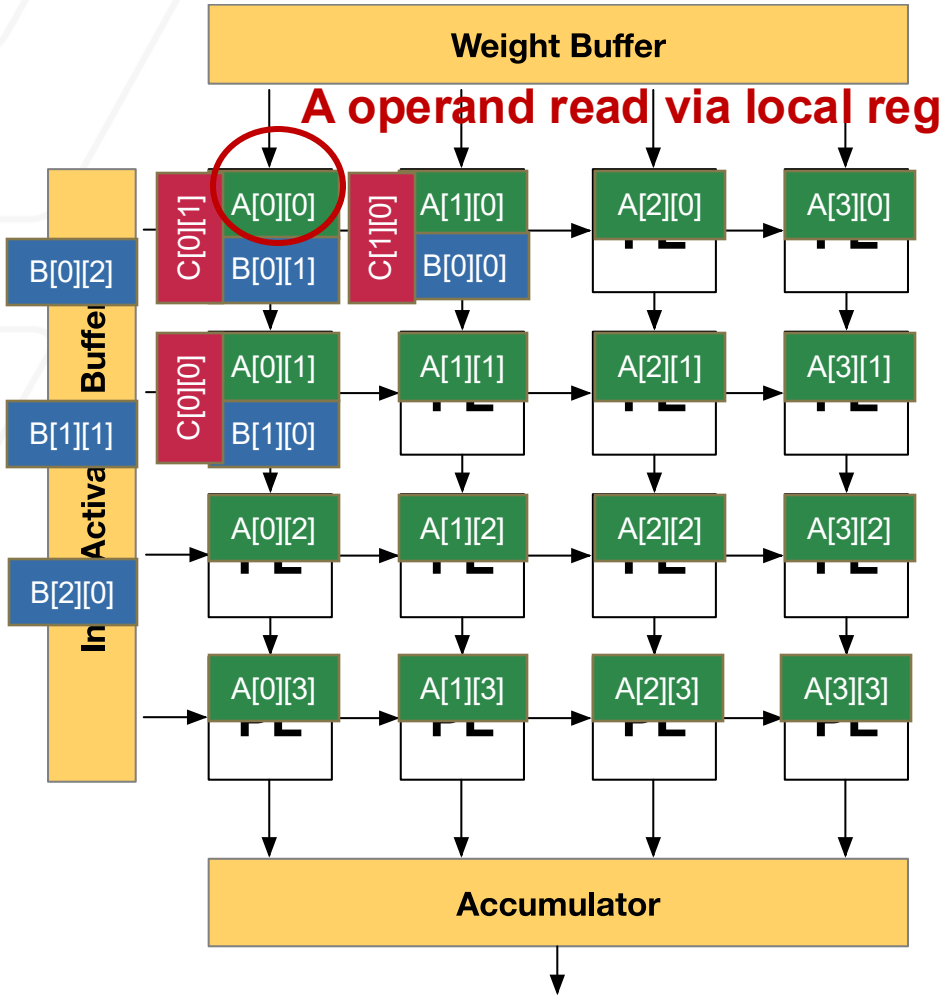
Assume following "mapping"

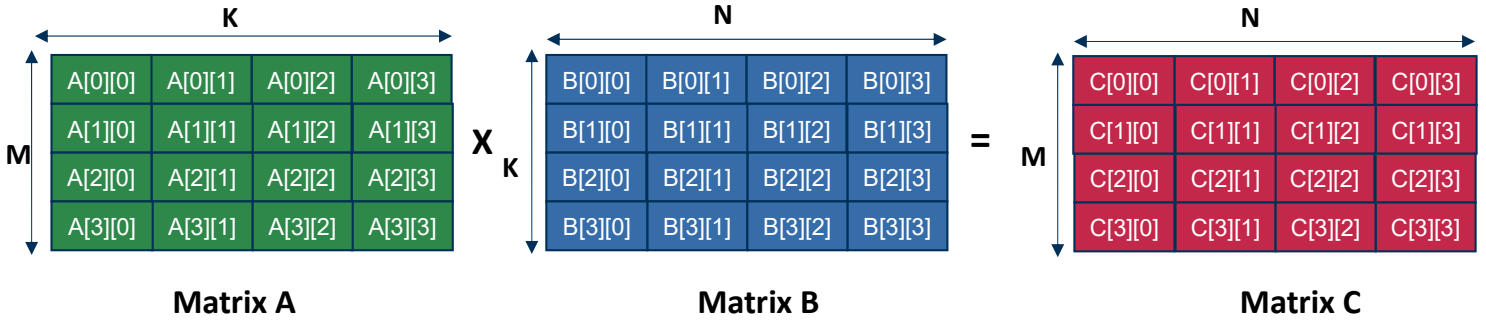
- Column: Reduced (*aka* contracted) Dimension (K)
- Row: Spatial Dimension (either M or N)

Note: alternate mapping styles possible

Matrix Multiplication on a Systolic Array

Phase 2:
Stream Activation Tensor






```
for m in range(4):
  for n in range(4):
    for k in range(4):
      C[m][n] += A[m][k] * B[k][n]
```

Assume following "mapping"

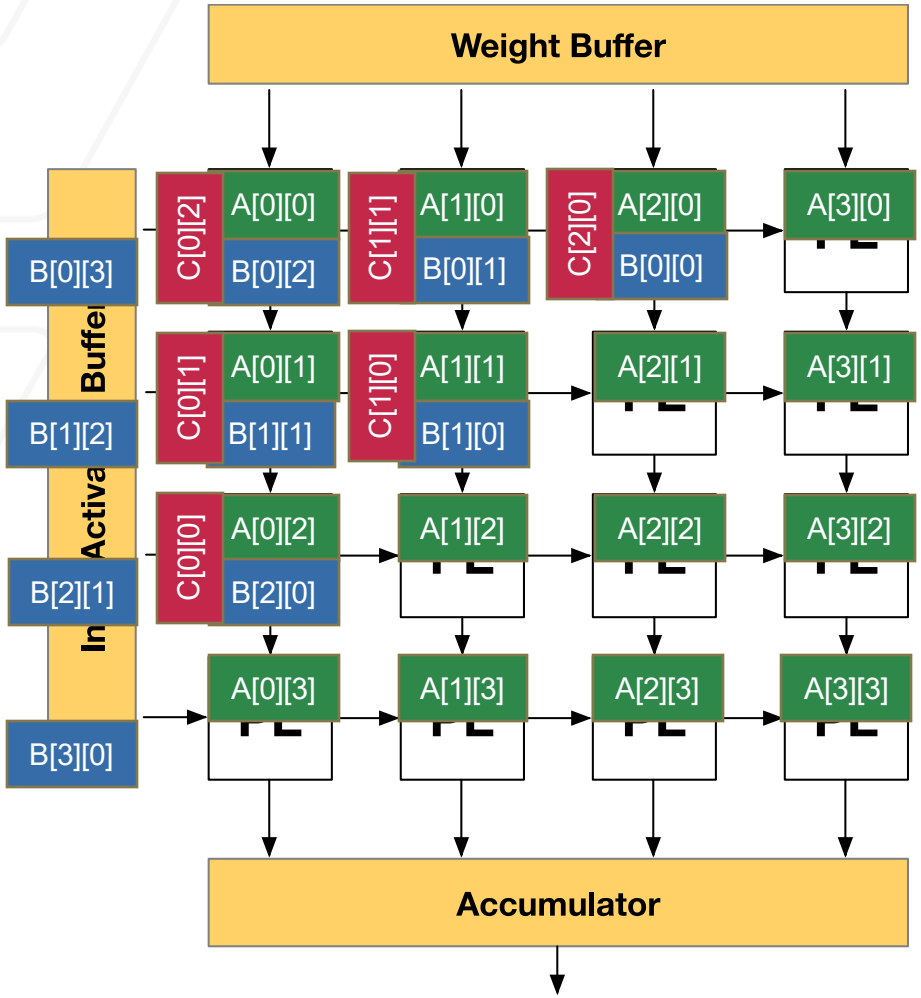
- Column: Reduced (*aka* contracted) Dimension (K)
- Row: Spatial Dimension (either M or N)

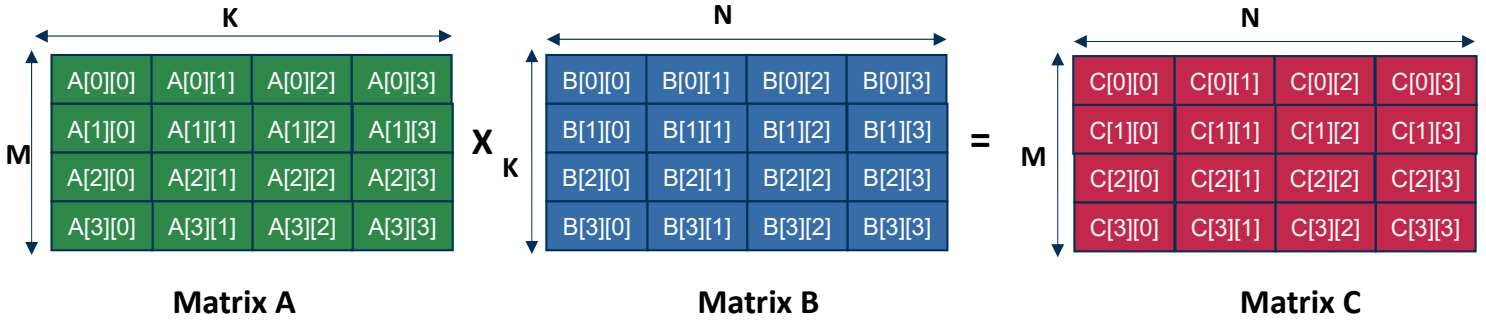
Note: alternate mapping styles possible

10

Matrix Multiplication on a Systolic Array

Phase 2:
Stream Activation Tensor






```
for m in range(4):
  for n in range(4):
    for k in range(4):
      C[m][n] += A[m][k] * B[k][n]
```

Assume following "mapping"

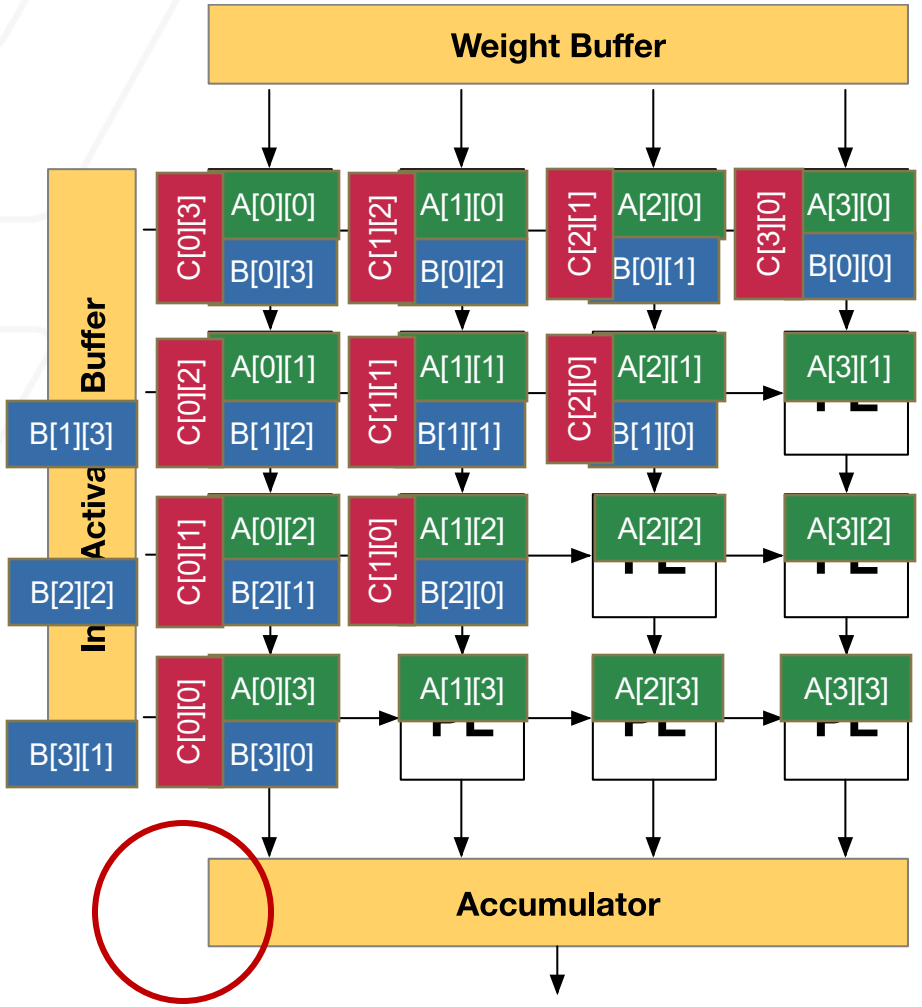
- Column: Reduced (*aka* contracted) Dimension (K)
- Row: Spatial Dimension (either M or N)

Note: alternate mapping styles possible

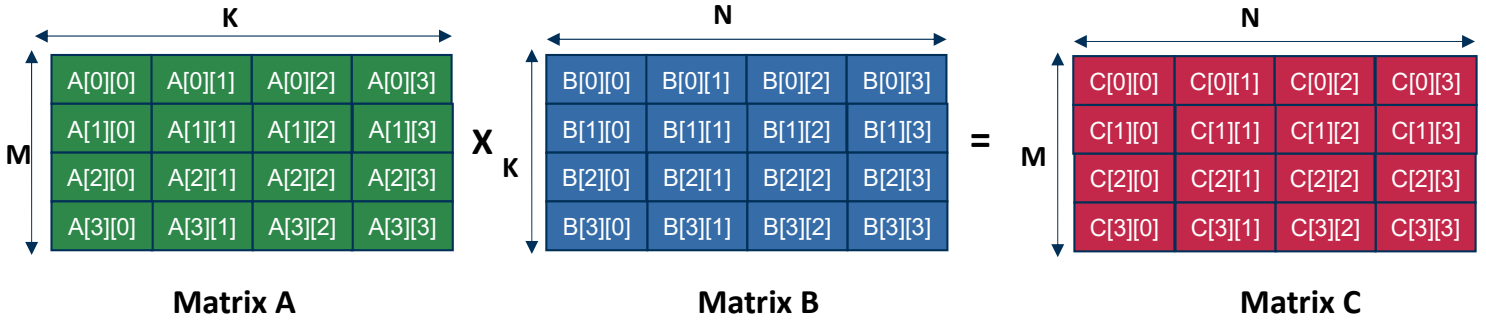
11

Matrix Multiplication on a Systolic Array

Phase 2:
Stream Activation Tensor



First output generated




```
for m in range(4):
  for n in range(4):
    for k in range(4):
      C[m][n] += A[m][k] * B[k][n]
```

Assume following "mapping"

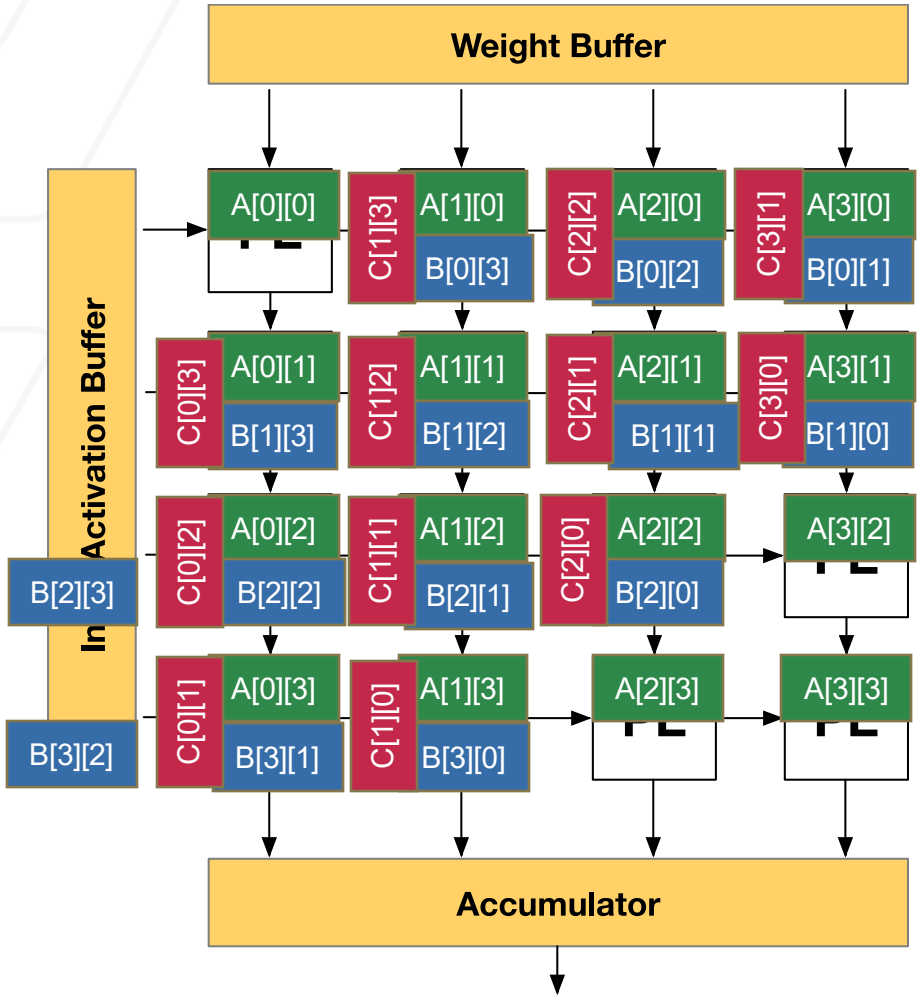
- Column: Reduced (*aka* contracted) Dimension (K)
- Row: Spatial Dimension (either M or N)

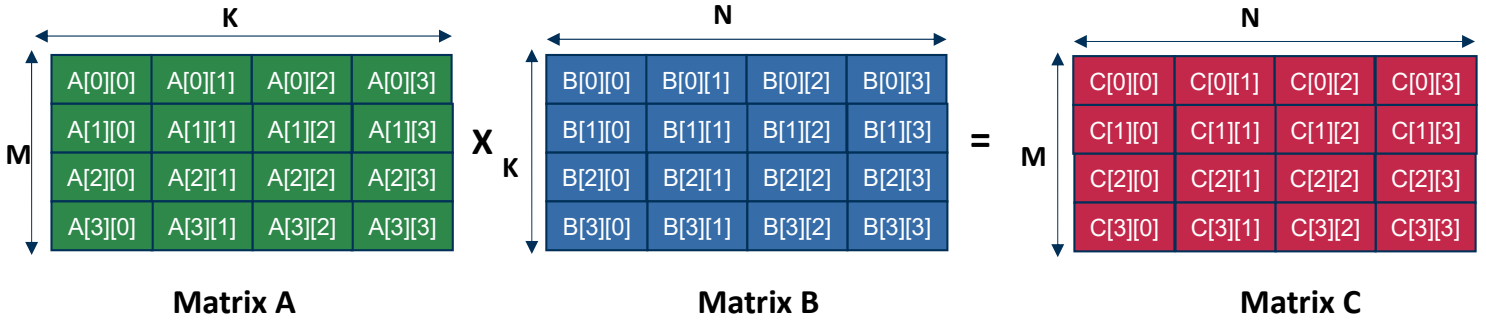
Note: alternate mapping styles possible

Georgia Tech

Matrix Multiplication on a Systolic Array

Phase 2:
Stream Activation Tensor






```
for m in range(4):
  for n in range(4):
    for k in range(4):
      C[m][n] += A[m][k] * B[k][n]
```

Assume following "mapping"

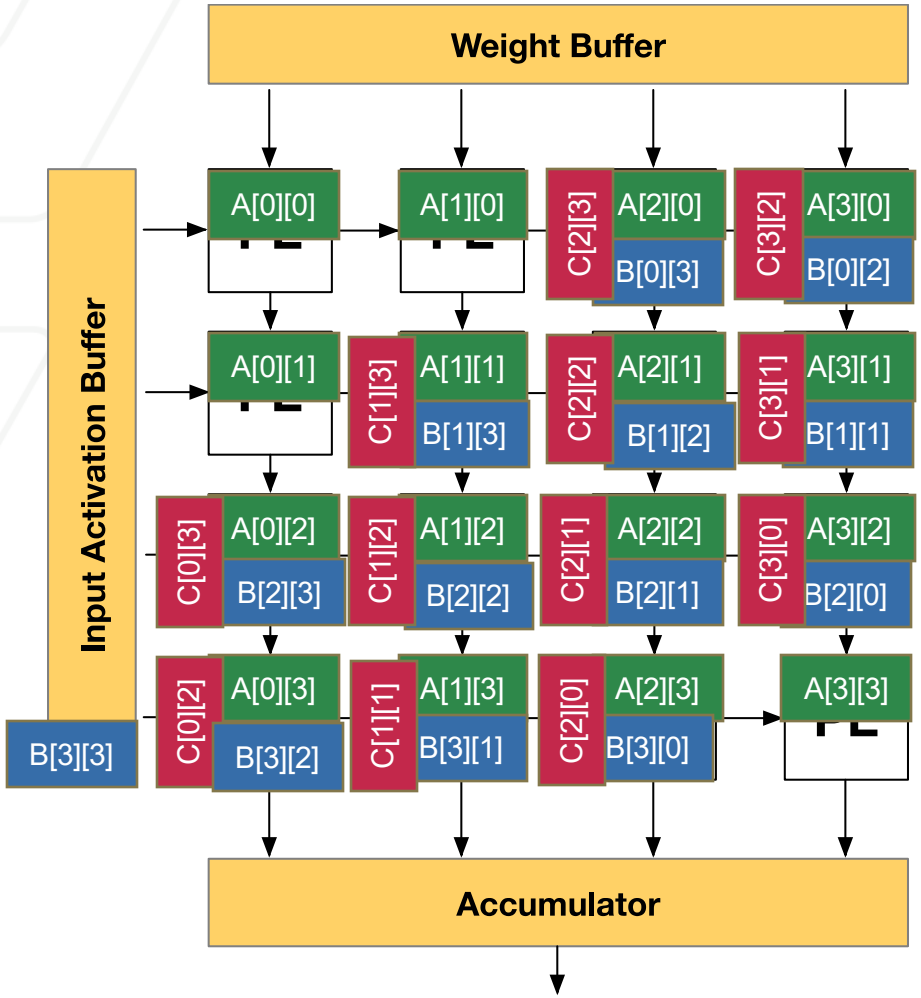
- Column: Reduced (*aka* contracted) Dimension (K)
- Row: Spatial Dimension (either M or N)

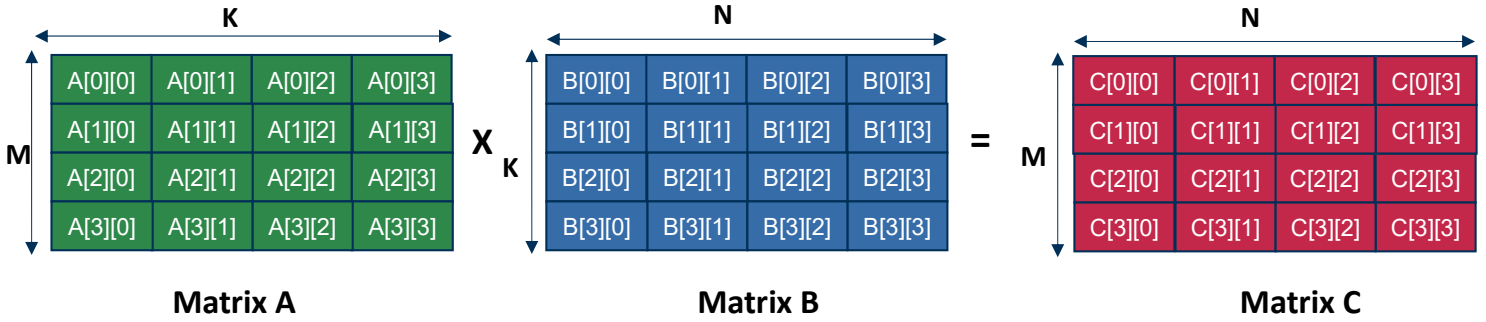
Note: alternate mapping styles possible

Georgia Tech

Matrix Multiplication on a Systolic Array

Phase 2:
Stream Activation Tensor






```
for m in range(4):
  for n in range(4):
    for k in range(4):
      C[m][n] += A[m][k] * B[k][n]
```

Assume following "mapping"

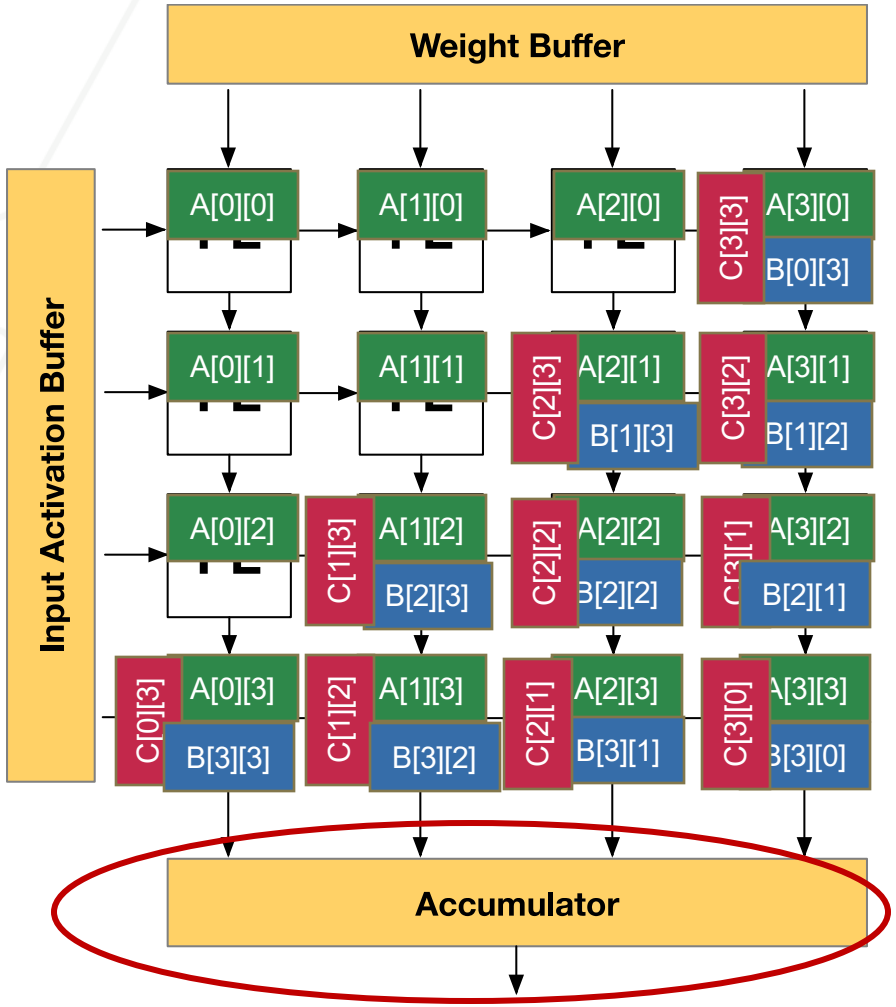
- Column: Reduced (*aka* contracted) Dimension (K)
- Row: Spatial Dimension (either M or N)

Note: alternate mapping styles possible

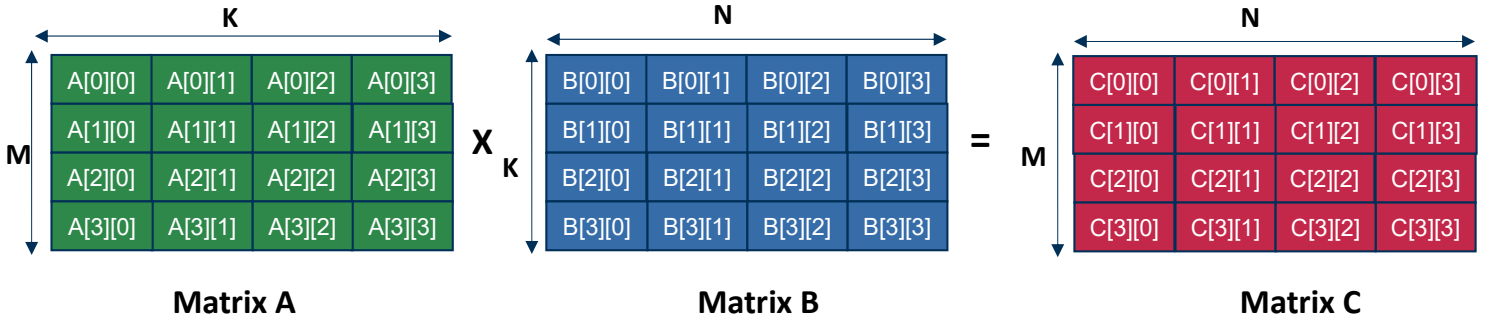
Georgia Tech

Matrix Multiplication on a Systolic Array

Phase 2:
Stream Activation Tensor



Four outputs every cycle in steady state




```
for m in range(4):
  for n in range(4):
    for k in range(4):
      C[m][n] += A[m][k] * B[k][n]
```

Assume following "mapping"

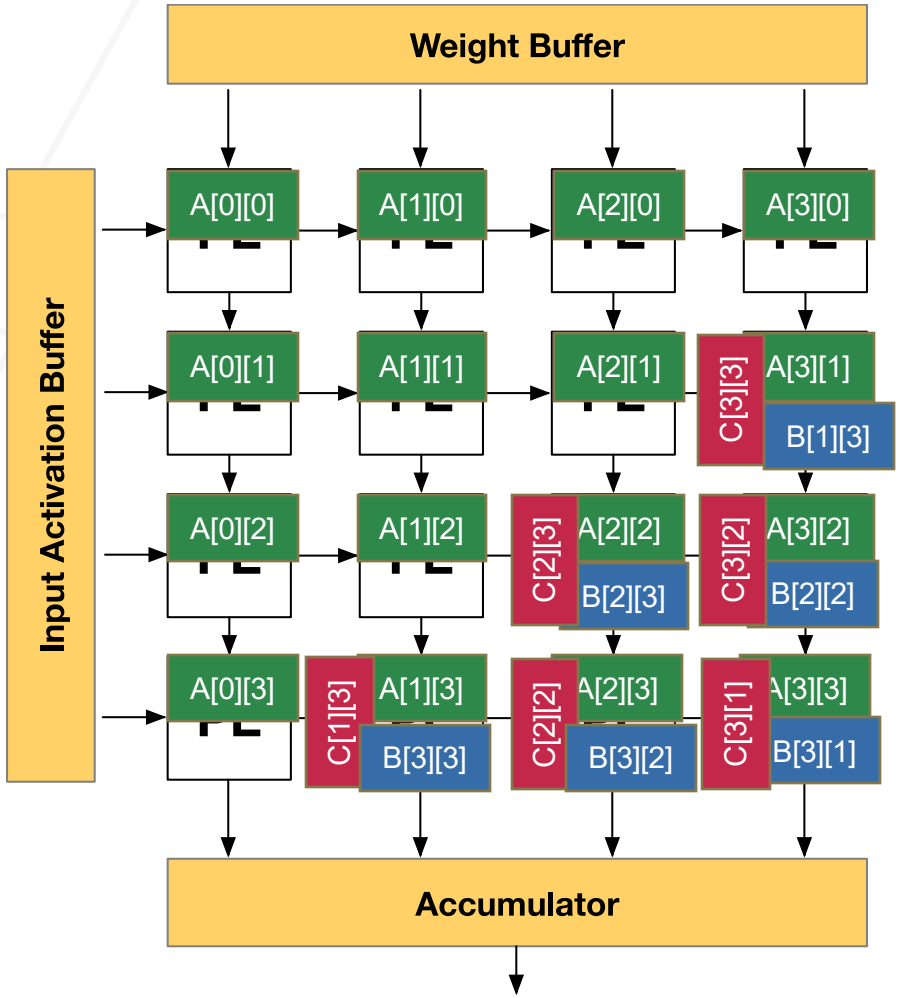
- Column: Reduced (*aka* contracted) Dimension (K)
- Row: Spatial Dimension (either M or N)

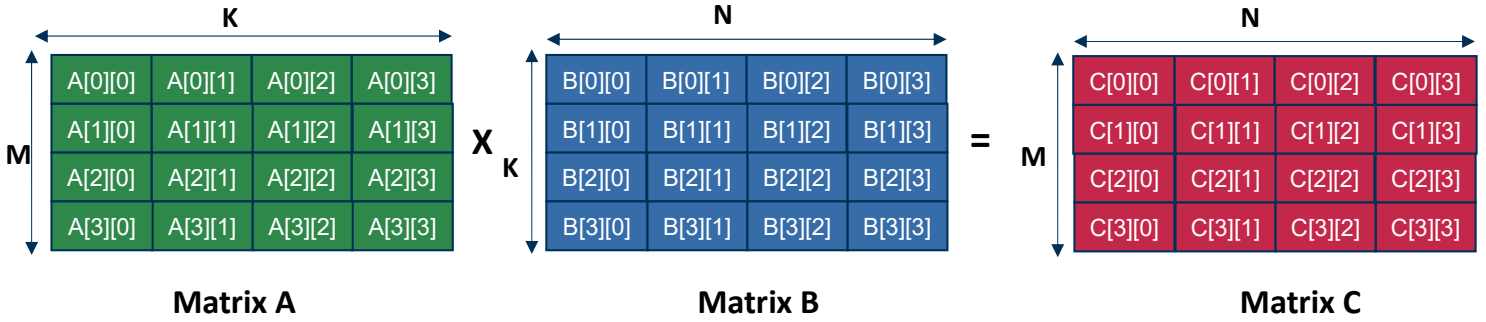
Note: alternate mapping styles possible

Georgia Tech

Matrix Multiplication on a Systolic Array

Phase 2:
Stream Activation Tensor






```
for m in range(4):
  for n in range(4):
    for k in range(4):
      C[m][n] += A[m][k] * B[k][n]
```

Assume following "mapping"

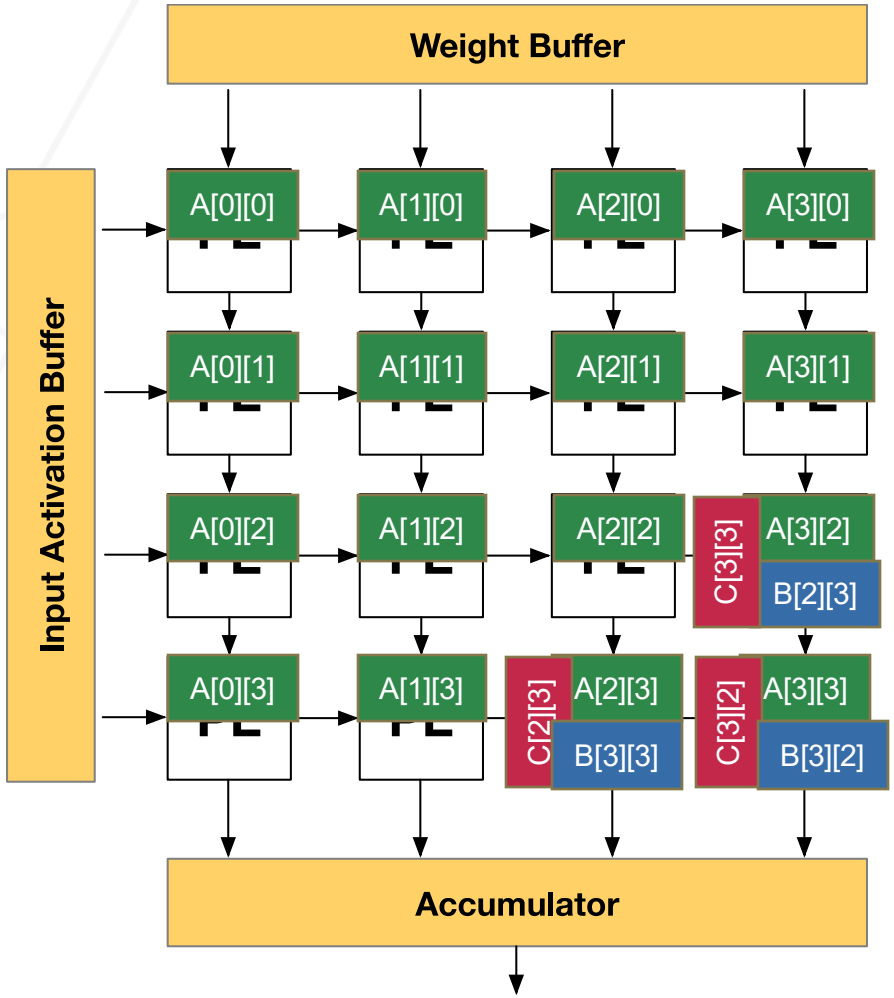
- Column: Reduced (*aka* contracted) Dimension (K)
- Row: Spatial Dimension (either M or N)

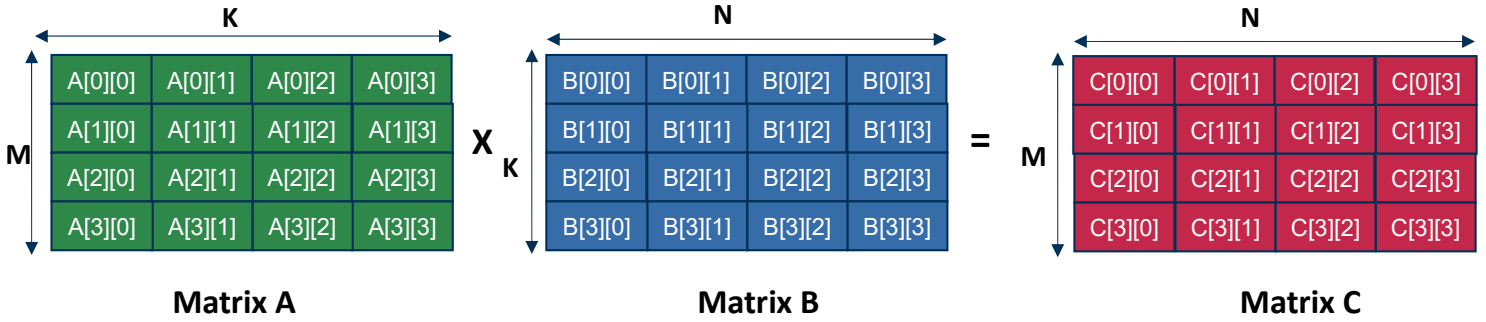
Note: alternate mapping styles possible

Georgia Tech

Matrix Multiplication on a Systolic Array

Phase 2:
Stream Activation Tensor






```
for m in range(4):
  for n in range(4):
    for k in range(4):
      C[m][n] += A[m][k] * B[k][n]
```

Assume following "mapping"

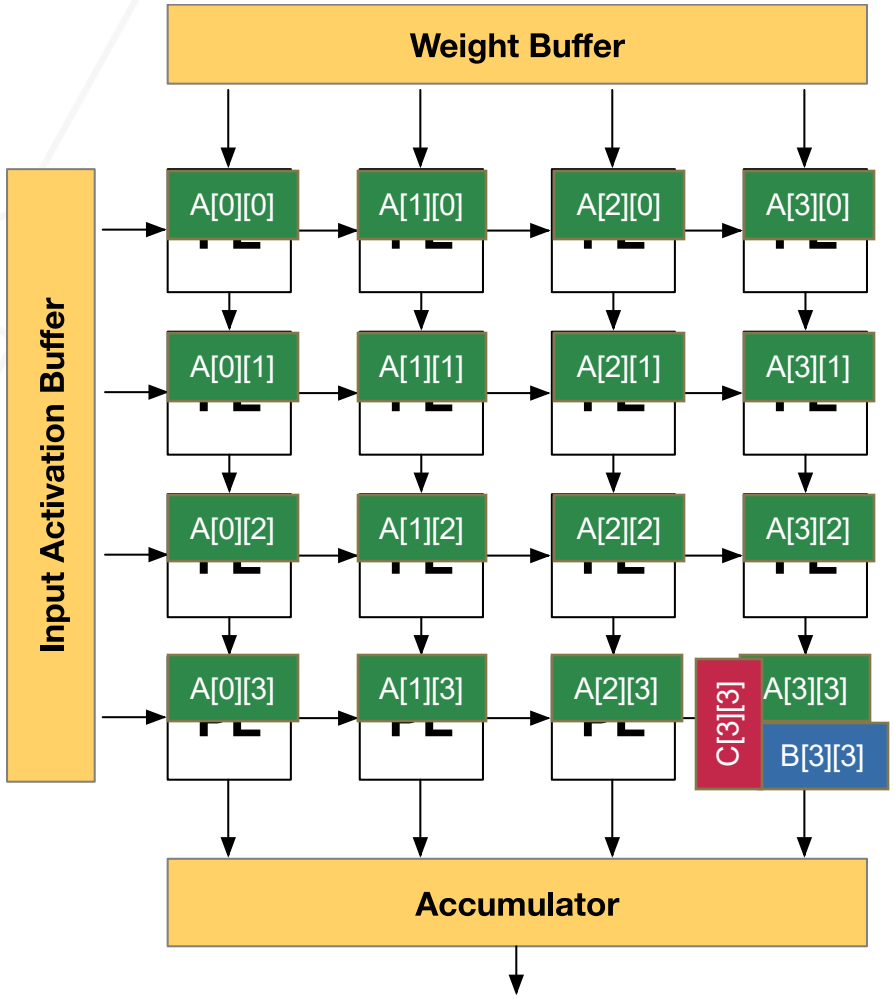
- Column: Reduced (*aka* contracted) Dimension (K)
- Row: Spatial Dimension (either M or N)

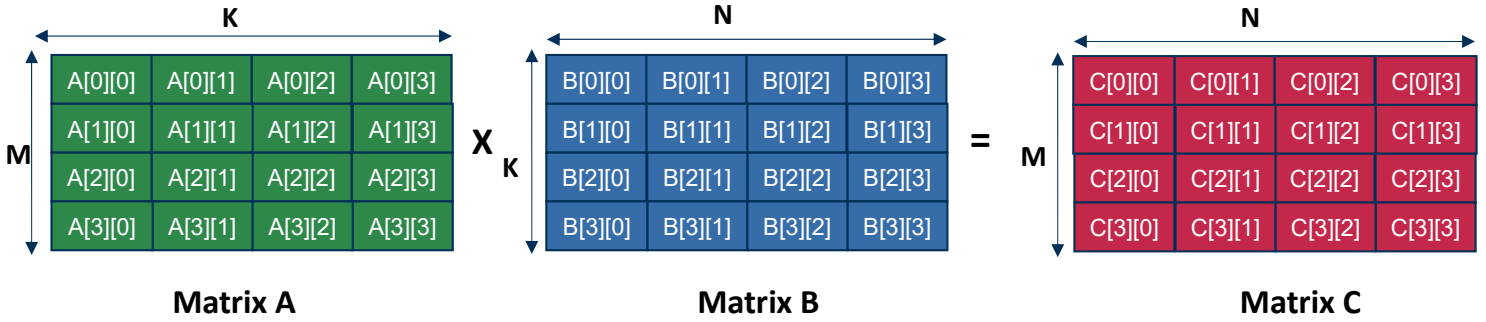
Note: alternate mapping styles possible

Georgia Tech

Matrix Multiplication on a Systolic Array

Phase 2:
Stream Activation Tensor






```
for m in range(4):
  for n in range(4):
    for k in range(4):
      C[m][n] += A[m][k] * B[k][n]
```

Assume following "mapping"

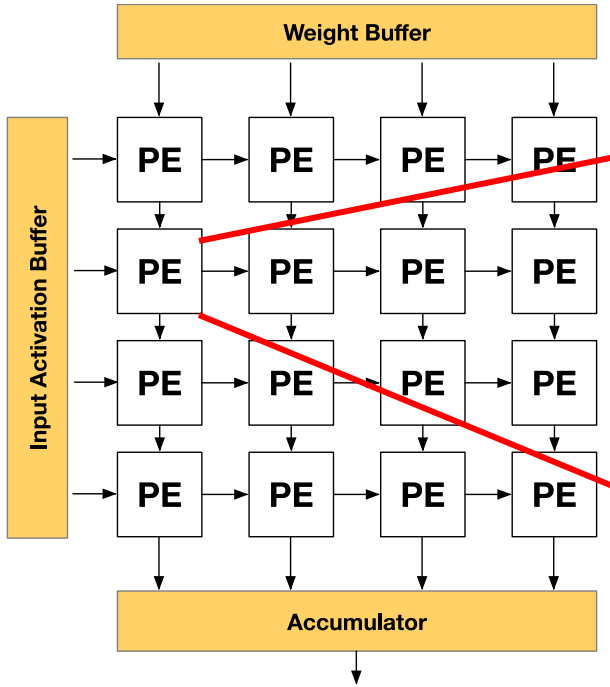
- Column: Reduced (*aka* contracted) Dimension (K)
- Row: Spatial Dimension (either M or N)

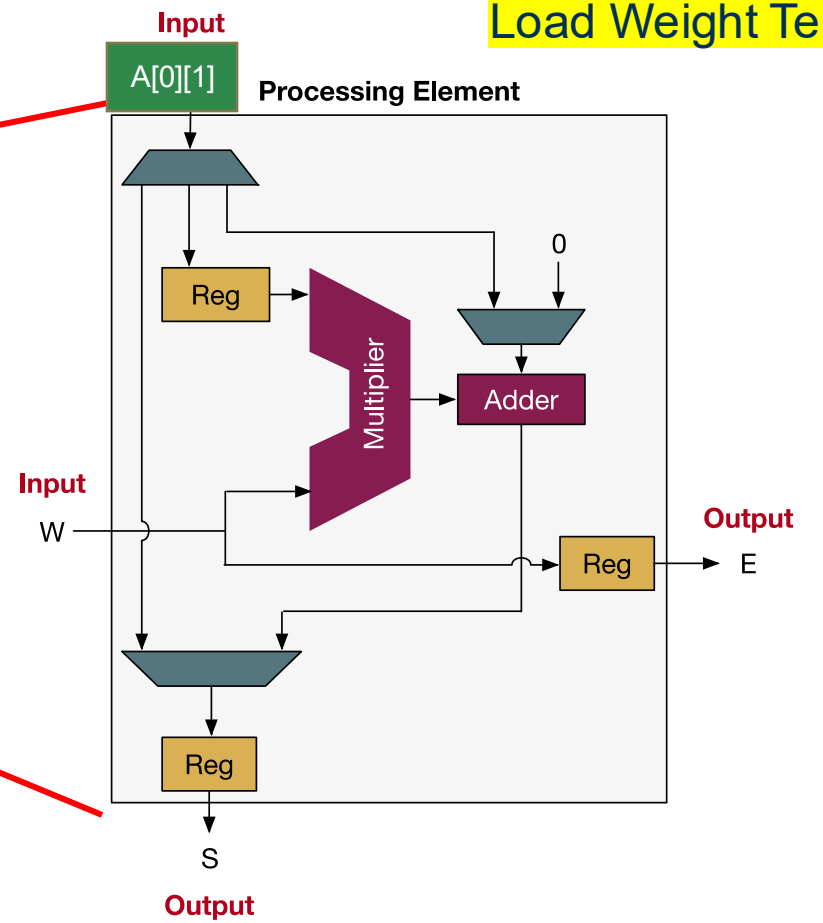
Note: alternate mapping styles possible

Georgia Tech

Recap: Zooming into PE

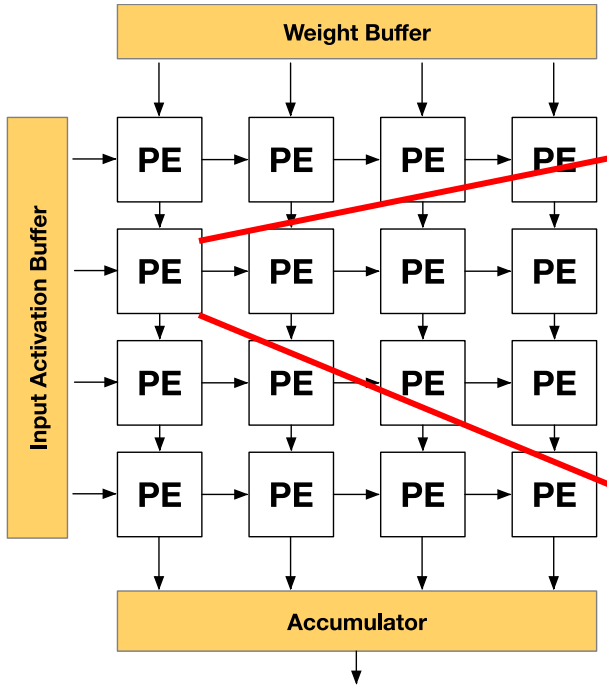
Phase 1:
Load Weight Tensor





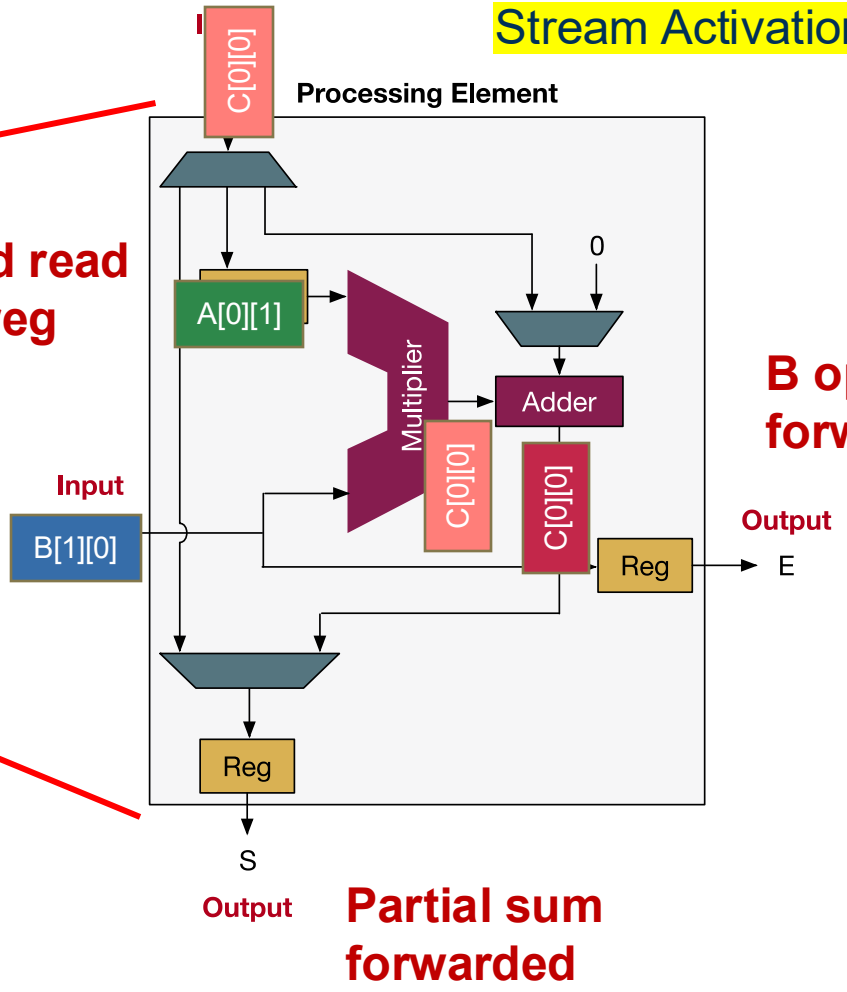
19

Recap: Zooming into PE



Phase 2:
Stream Activation Tensor

A operand read via local reg



B operand forwarded

Partial sum forwarded

Georgia Tech

20