Session 5A: Frameworks for
deep learning — Layering the ML cake.

ASPLOS'20, March 16–20, 2020, Lausann

# CS 6230
## Term Project
### Due 11:59pm, Wednesday, 12/21/2022

The term project can be done either alone or in teams of two. When grading, some consideration will be given to whether it was a solo effort or a two-person effort, with higher expectations for projects don...

For the proje... mat...
you have a ... **C=**
computation... e CP
GPUs. A p... n ex
assignments ... *nera*
exercises w... nly
(equal value... on) a
a single prob... of de
numerical lib... of c
the specific ... ng th
known a prio... Thus, adaptivity with respect to the problem size parameters in the call can be beneficial. The adaptivity may involve dynamically choosing parameters such as the sizes, shapes/sized of thread blocks, or even different code versions, e.g., representing different loop permutations and/or degrees of loop unrolling. If multiple code versions are used, an automatic dynamic selection mechanism must be implemented that chooses one of the versions based on problem size parameters.

The two code versions to be optimized:

**1) Transposed-Nontransposed Matrix-Matrix Multiplication (C=AᵀB):**

```
for (i=0;i<Ni;i++)
  for (k=0;k<Nk;k++)
    for (j=0;j<Nj;j++)
// C[i][j] += A[k][i]*B[k][j];
      C[i*Nj+j] += A[k*Ni+i]*B[k*Nj+j];
```

**2) Transposed-Transposed Matrix-Matrix Multiplication (C=AᵀBᵀ):**

```
for (i=0;i<Ni;i++)
  for (k=0;k<Nk;k++)
    for (j=0;j<Nj;j++)
// C[i][j] += A[k][i]*B[j][k];
      C[i*Nj+j] += A[k*Ni+i]*B[j*Nk+k];
```

Like the programming assignments, template codes will be provided, along with a number of test cases with varying sizes of the matrices. Here are examples of problem-size variation for test cases:

| Ni | Nj | Nk |
|---|---|---|
| 8*1024 | 8*1024 | 16 |
| 4096 | 4096 | 64 |
| 2048 | 2048 | 256 |
| 1024 | 1024 | 1024 |
| 256 | 256 | 16*1024 |
| 64 | 64 | 256*1024 |
| 16 | 16 | 4*1024*1024 |

| Ni | Nj | Nk |
|---|---|---|
| 9*999 | 9*999 | 37 |
| 3*999 | 3*999 | 111 |
| 999 | 999 | 999 |
| 333 | 333 | 9*999 |
| 111 | 111 | 81*999 |
| 37 | 37 | 729*999 |

---



Figure 7. Validation of the analytical model against post-synthesis results and previous published mode...

**Table 3.** Energy per 16-bit access with various register file (RF) and SRAM sizes, and for a MAC operation, one hop

| RF-Size | Energy (pJ) |
|---|---|
| 16 B | 0.03 |
| 32 B | 0.06 |
| 64 B | 0.12 |
| 128 B | 0.24 |
| 256 B | 0.48 |
| 512 B | 0.96 |

| SRAM Size | Energy (pJ) |
|---|---|
| 16 KB | 3 |
| 32 KB | 6 |
| 64 KB | 9 |
| 128 KB | 13.5 |
| 256 KB | 20.25 |
| 512 KB | 30.375 |
| DRAM | 200 |

| | Energy (pJ) |
|---|---|
| MAC | 0.075 |
| Hop | 0.035 |

**Table 4.** ASIC designs for model valida...

| Name | Dataflow | PE Array | RF | S |
|---|---|---|---|---|
| OS4 | X | 1D, 4 | 32 B | 3 |
| OS8 | X | 1D, 8 | 64 B | 6 |
| WS16 | C\|K | 2D, 4×4 | 64 B | 3 |

**Table 5.** Energy per op, in pJ, for various ops in 45 nm. The second column in each group shows energy of multiple versus a single add operation.

| Operation | 16 bit (int) | | 64 bit (dp) | |
|---|---|---|---|---|
| | E/op | vs. Add | E/op | vs. Add |
| Add | 0.18 | 1.0x | 5 | 1x |
| Multiply | 0.62 | 3.4x | 20 | 4x |
| 16-word RF | 0.12 | 0.7x | 0.34 | 0.07x |
| 4K-word SRAM | 8 | 44x | 26 | 5.2x |
| 32K-word SRAM | 11 | 61x | 47 | 9.4x |
| DRAM | 640 | 3556x | 2560 | 512x |