DRAM Access Reduction by Node Fusion with TVM

Chia-Wei Chang, Jing-Jia Liou,

Chih-Tsun Huang, Wei-Chung Hsu & Juin-Ming Lu

National Tsing Hua University & Industrial Technology Research Institute

Dec 5th, 2019

國立清華大學
NATIONAL TSING HUA UNIVERSITY

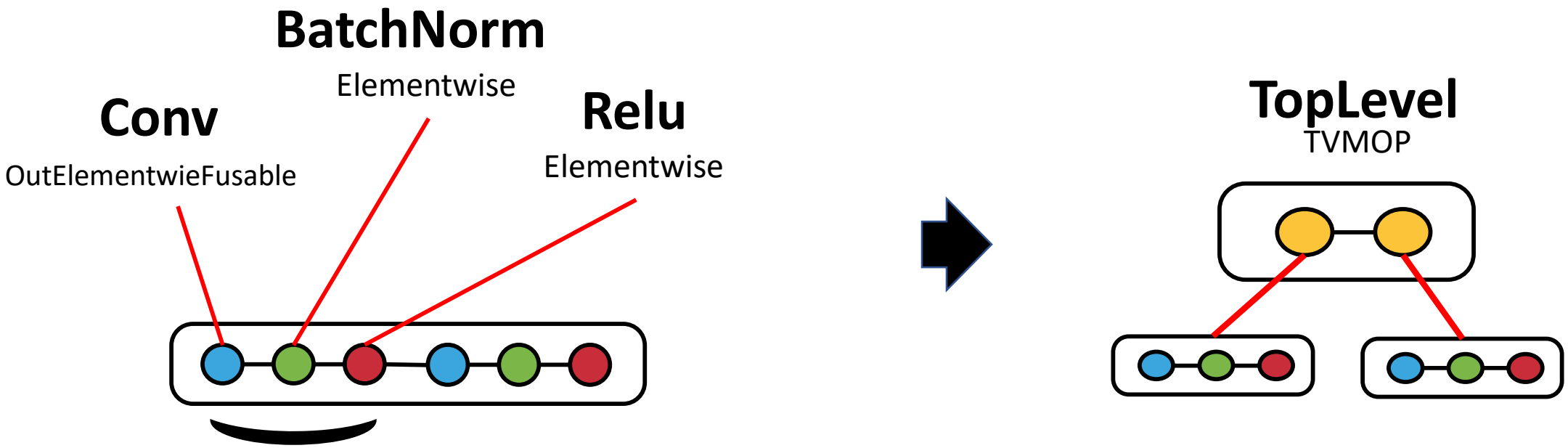
工業技術研究院
Industrial Technology
Research Institute

DRAM Access Consumes More Energy

- Energy efficiency is the key to DNN computation
  - Hardware accelerators
- DRAM consumes 50-100x more energy per byte than SRAM
- Node fusion is used to save DRAM accesses

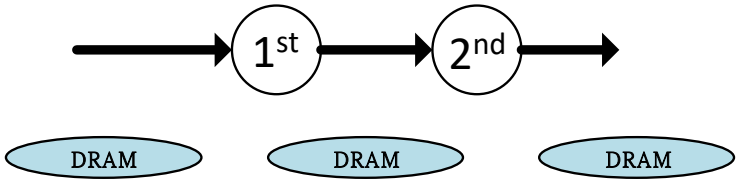
	DRAM	SRAM	Register
Energy	250x	4x	1x

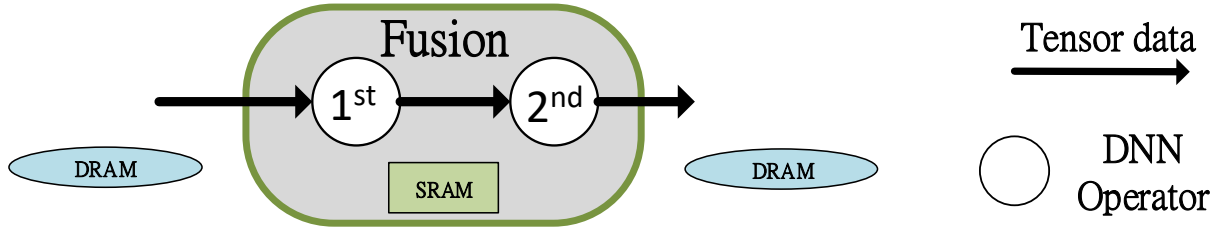
TVM only Fuses Elementwise OP



- Currently, TVM only supports fusion of elementwise OP into Conv
- Each OP has an attribute to indicate whether to fuse
- Generate TVMOP, which includes nodes to share data in SRAM

Our Node Fusion Merges Multiple Convs



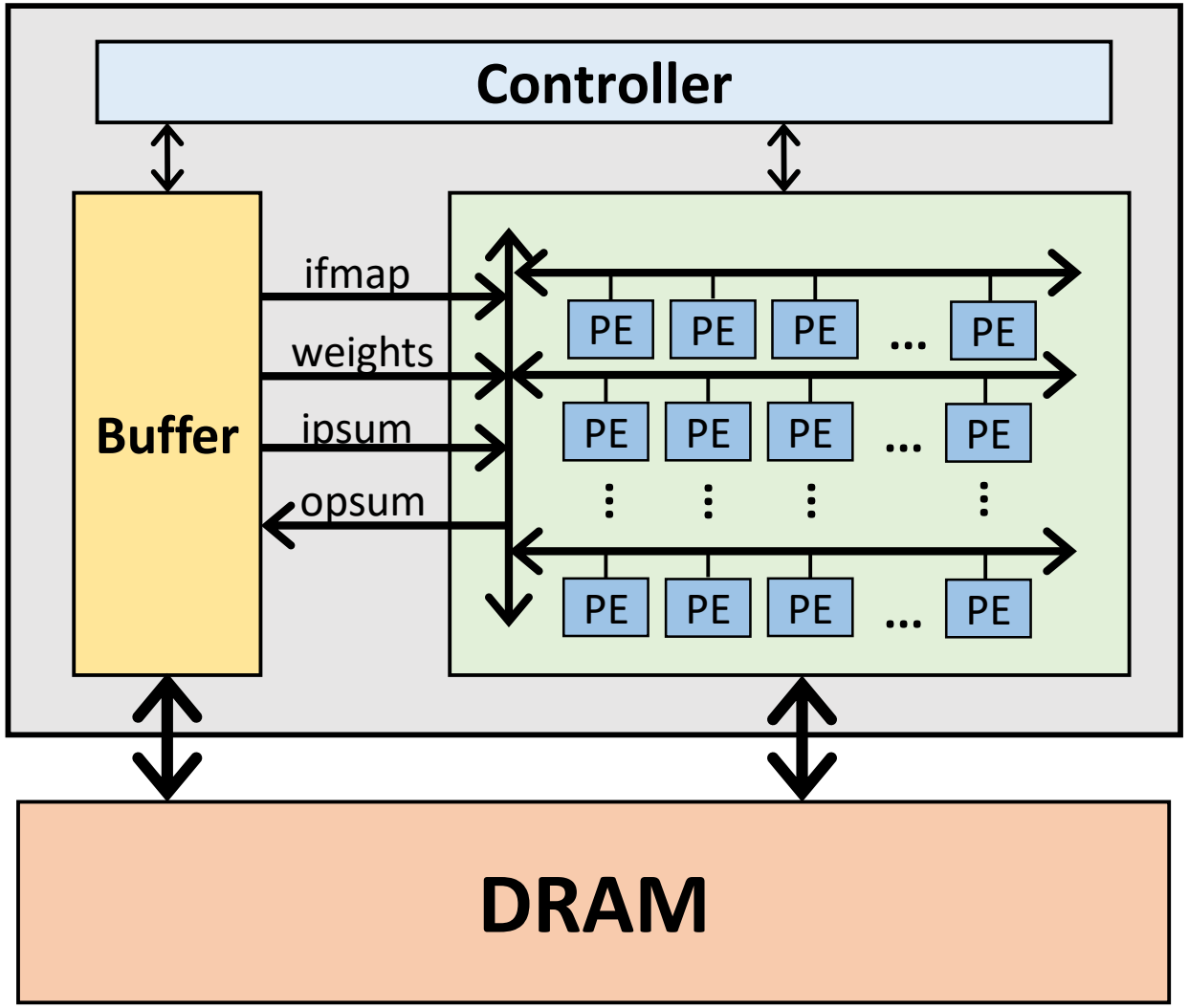



```
for (n=0; n<N; n++)                    # 1st Conv
for (k=0; k<C1; k++)
for (y=0; y<H1; y++)
for (x=0; x<W1; x++)
   for (c=0; c<C0; c++)
   for (r=0; r<R1; r++)
   for (s=0; s<S1; s++)
      O1[n][k][y][x] += W1[k][c][r][s] * I[n][c][y+r][x+s]

for (n=0; n<N; n++)                    # 2nd Conv
for (k=0; k<C2; k++)
for (y=0; y<H2; y++)
for (x=0; x<W2; x++)
   for (c=0; c<C1; c++)
   for (r=0; r<R2; r++)
   for (s=0; s<S2; s++)
      O2[n][k][y][x] += W2[k][c][r][s] * O1[n][c][y+r][x+s]
```

```
for (n=0; n<N; n++)
for (k=0; k<C2; k++)
for (y=0; y<H2; y++)
for (x=0; x<W2; x++)
   int sram[C1][R2][S2]        # Internal SRAM buffer
   for (c=0; c<C1; c1++)
   for (r=0; r<R2; r++)
   for (s=0; s<S2; s++)
     for (c2=0; c2<C0; c++)
     for (r2=0; r2<R1; r++)
     for (s2=0; s2<S1; s++)
       sram[c][r][s] += W1[c][c2][r2][s2] * I[n][c2][y+r+r2][x+s+s2]

   for (c=0; c<C1; c++)
   for (r=0; r<R2; r++)
   for (s=0; s<S2; s++)
      O[n][k][y][x] += W2[k][c][r][s] * sram[c][r][s]
```

Experiment Settings: Hardware

- Eyeriss-like architecture
- 256MB DRAM
- 108KB SRAM
- 12x14 PE
- Runs AlexNet
- Due to hardware limitation, only Conv is evaluated



Experimental Results