# swDNN:A Library for Accelerating Deep Learning Applications on Sunway TaihuLight

Jiarui Fang, Haohuan Fu, Wenlai Zhao, Bingwei Chen, Weijie Zheng, Guangwen Yang,

[1]Department of Computer Science and Technology, Tsinghua University
[2]Ministry of Education Key Lab. for Earth System Modeling, Department of Earth System Science, Tsinghua University [3]National Supercomputing Center in Wuxi

2017 IEEE International Parallel and Distributed Processing Symposium
Presented by Ching-Yuan, Tsai

# Outline

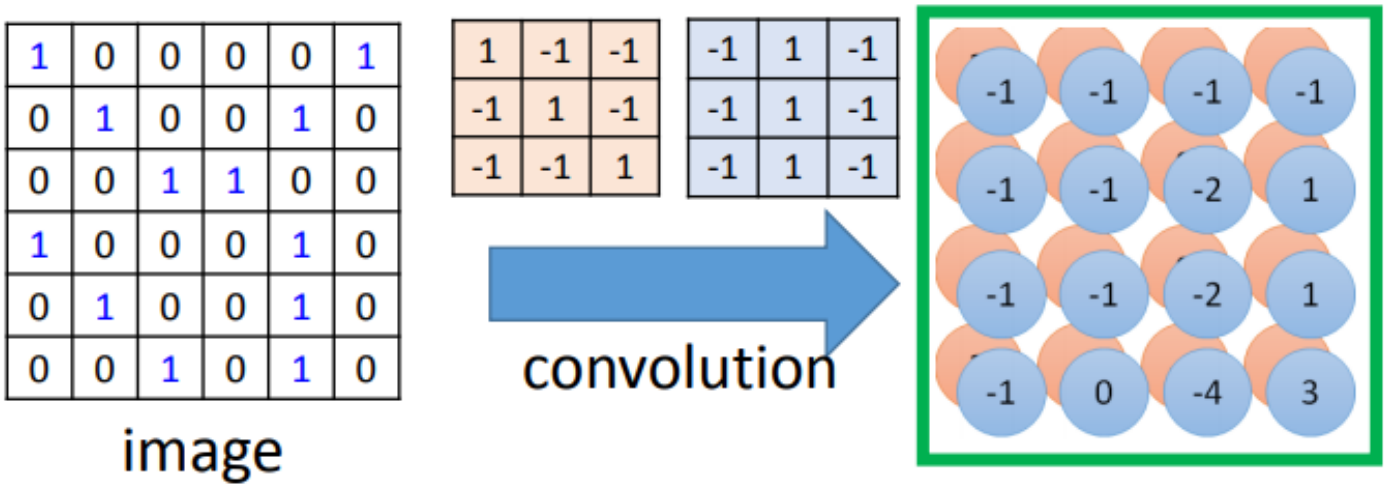
# Sunway TaihuLight

- A supercomputer that ranks the first in the world.
- Over 100 Pflops computing capacity.
- Sunway TaihuLight is power by a new SW26010 many-core processor.
- Not only the fastest but also the greenest supercomputer in the world.

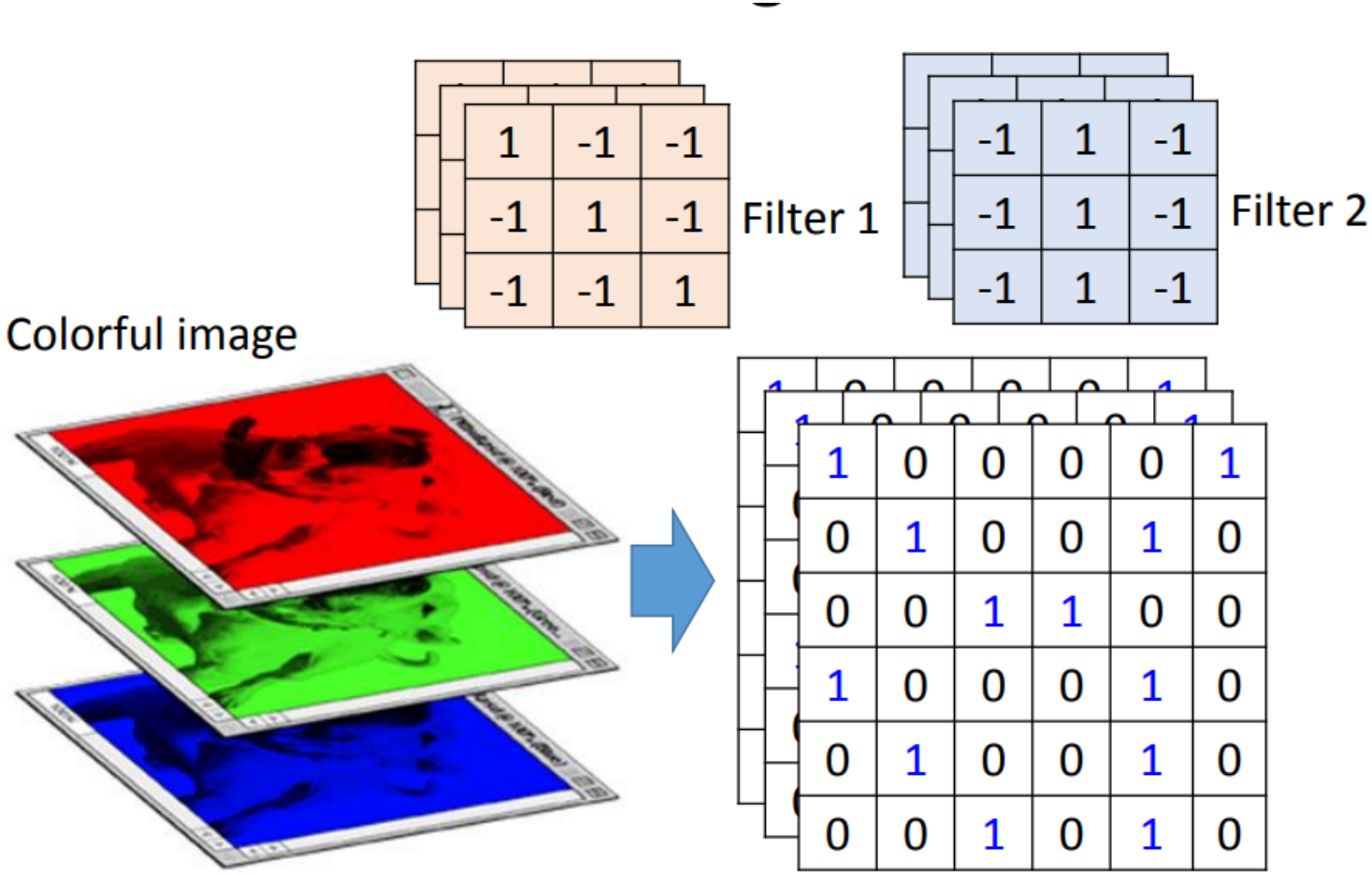
# SW26010

- Peak double-precision performance of 3.06 Tflops.
- 300 watts power consumption.
- Combine both management cores and computing core clusters on a core group.
- Support a user-controlled fast buffer for each computing cores.
- Rigister communication between computing cores.
- Each computing core consists of two execution pipelines.

# Convolution layer



# Convolution layer



# Peseudo code of a convolutional layer

Table I: Parameters of convolutional layers

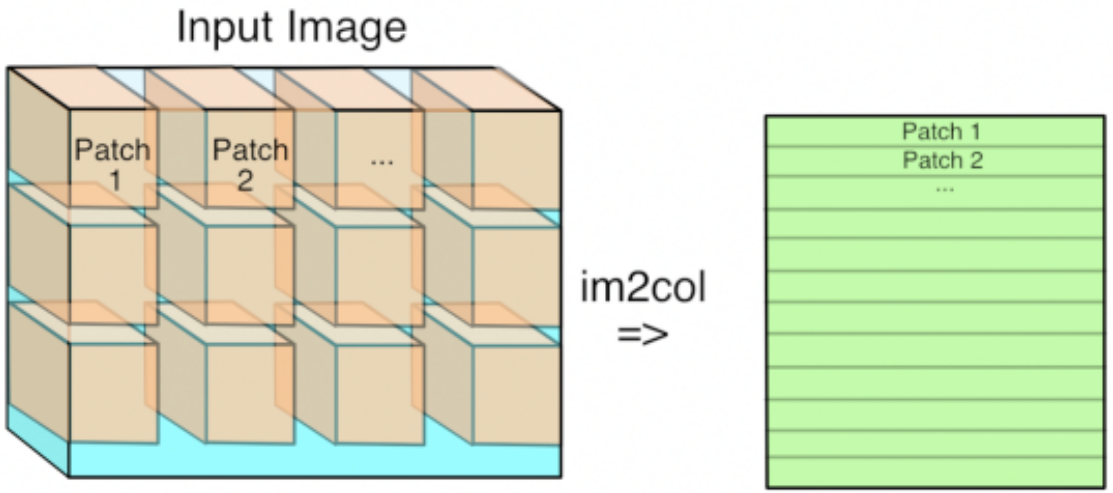
| Parameter | Meaning |
|---|---|
| $N_i$ | Number of input feature maps |
| $N_o$ | Number of output feature maps |
| $R_i$ | Height of input image |
| $C_i$ | Width of input image |
| $R_o$ | Height of output image |
| $C_o$ | Width of output image |
| $K_r$ | Height of filter kernel |
| $K_c$ | Width of filter kernel |

# Pseudo code of a convolutional layer

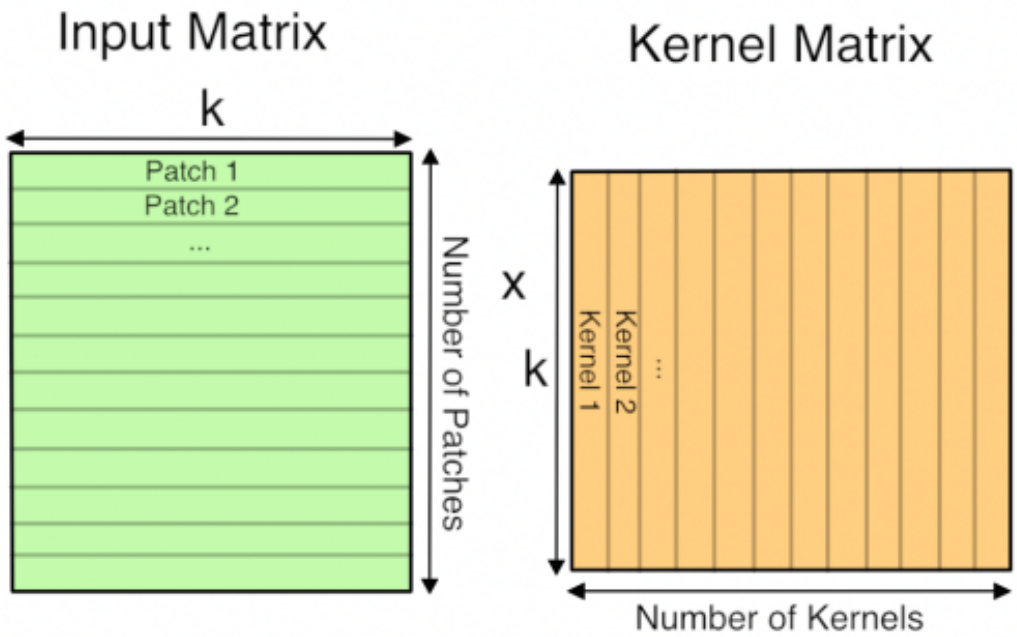
for $cB = 0$ to B
  for $cR_o = 0$ to $R_o$
    for $cC_o = 0$ to $C_o$
      for $cN_o = 0$ to $N_o$
        for $cK_r = 0$ to $K_r$
          for $cK_c = 0$ to $K_c$
            for $cN_i = 0$ to $N_i$
              out$[cB][cR_o][cC_o][cN_o]$ +=

in$[cB][cR_o + cK_r][cC_o + cK_c][cN_i]$*filter$[cN_o][cK_r][cK_c][cN_i]$

# General Matrix-Multiplication(GEMM)



# GEMM


Input Matrix
k
Patch 1
Patch 2
...
Number of Patches
Kernel Matrix
x
k
Kernel 1
Kernel 2
...
Number of Kernels

# Outline

# Architecture

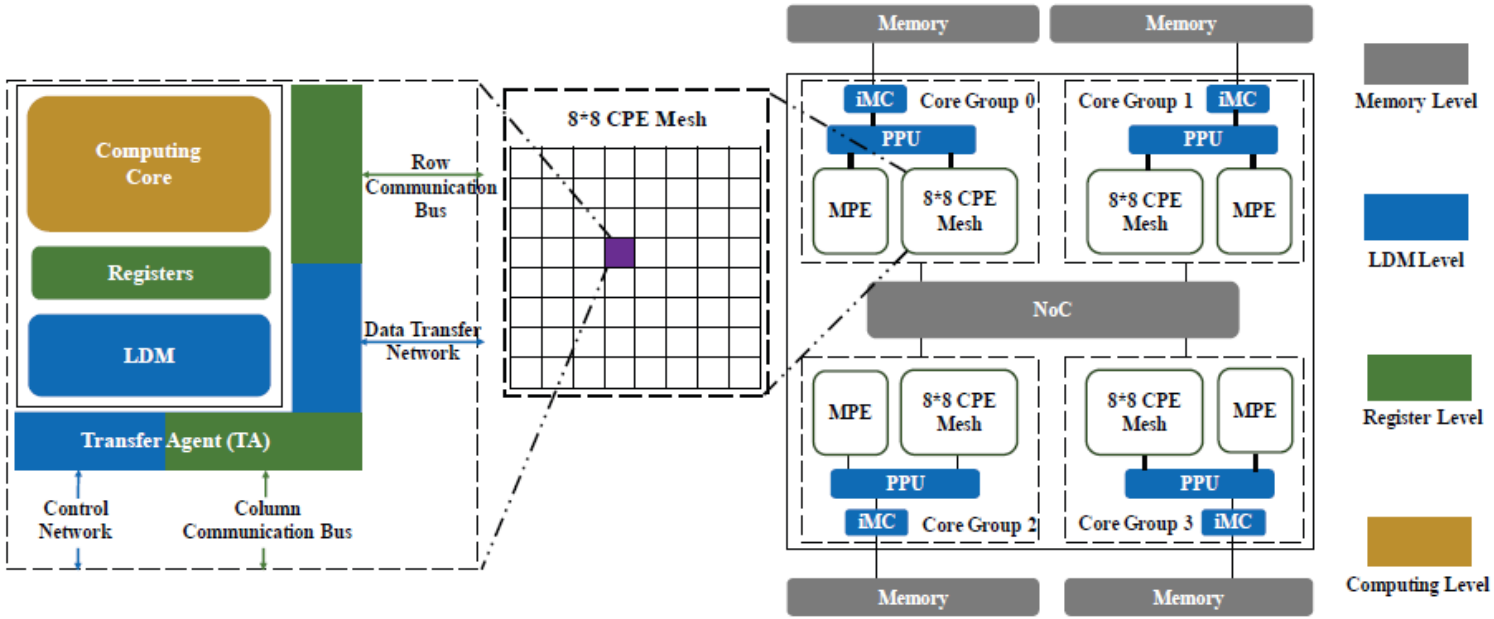



Figure 1: The general architecture of the SW26010 many-core processor.

# Unique features

- Each CG has an MPE.
- Users can explicitly set the sizeof each CG's private memory space, and the size of shared memory space.
- Support a 64kb user-controlled fast buffer for each computing cores.
- Rigister communication between computing cores.
- Each computing core consists of two execution pipelines.

# Challenges

- Low memory bandwidth.
  - SW26010:36 GB/s for each CG, 144 GB/S for entire processor.
  - NVIDIA K80GPU: 480 GB/S for entire processor.
- CPEs do not have a shared buffer to rely on a fine-grained data sharing scheme.

# Outline

# Register Communication

- Motivation
  - Fine-grained.
  - No shared memory in a CG.
- Architecture
  - 8 row communication buses.
  - 8 col communication buses.
  - Broadcast.

# $W * D_i = D_o$

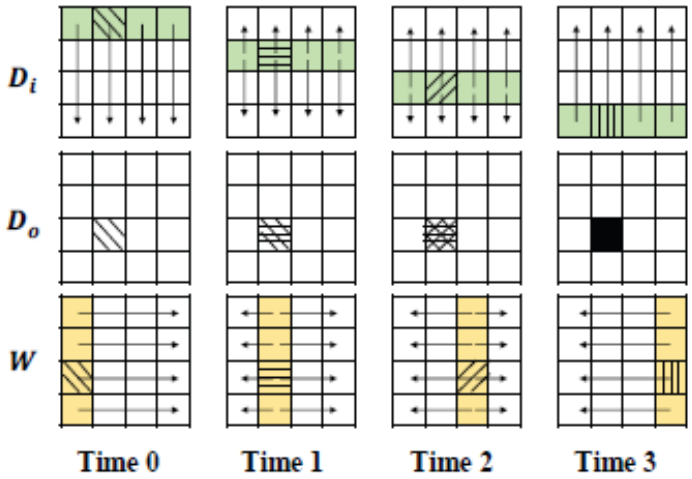



Figure 3: Schematic of register communication on CPEs for matrix multiplication.

# Outline

# P1 and P0

- Each CPE consists of two execution pipelines.
  - P0 : floating-pointer, vector operations.
  - P1 : Control transfer, load/store and register communication operations.
- The two execution pipelines share an Instruction Decoder and an instruction queue.
- In each cycle, two instructions in the front of the queue are issued into two pipelines.

# Instructions scheduling rules

- Both instructions have no conflicts with the unifinished instructions issued before.
- The two insturctions have no RAW or WAW conflicts.
- The two instructions can be handled by two execution pipelines separately.

# Outline

# Experiment

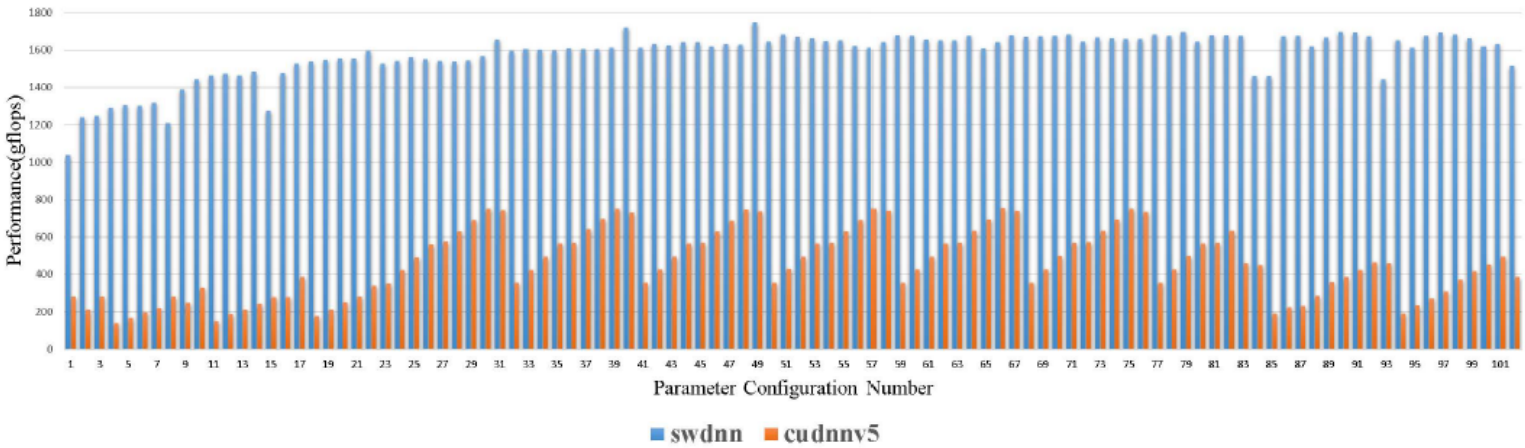



Figure 7: Double-precision performance results of our convolution kernels with different $(N_i, N_o)$ ranging from $(64, 64)$ to $(384, 384)$, compared with the K40m GPU results with cuDNNv5. ($B = 128$, output image $=64 \times 64$, filter $= 3 \times 3$)