

# Power Consumption of GPU/AI L1/L2 Cache

Chih-Cheng Hsiao 2023/08/16

# Overview

- The new invention can direct use on CPU L1/L2 cache, but data width of GPU/AI L1/L2 cache is too large, so there need a new partition to reduce GPU/AI L1/L2 cache power consumption.
- With new partition, GPU/AI L1 cache power consumption reduces to only 30% of original one, GPU/AI L2 cache power consumption reduces to only 15 % of original one.
- The new partition can suit more large data width, so data width can up to 512 bytes => provide more instruction/data to calculation unit.

Thus the performance will better and better.

# Original L1 Cache and New Partition

- The original GPU/AI L1 cache is shown in Fig. 1. Due to the data width is very large(256 bytes), the power consumption is very high.
- The new invention is shown in Fig. 2. The SRAM is divided to 16 small parts. The size of each small part is shown in Fig. 3. And Fig. 4 is the new invention applied to small part. Bit lines in each byte is shown in Fig. 5. Bit line arrangement is shown in Fig. 6.

# Original L1 Cache

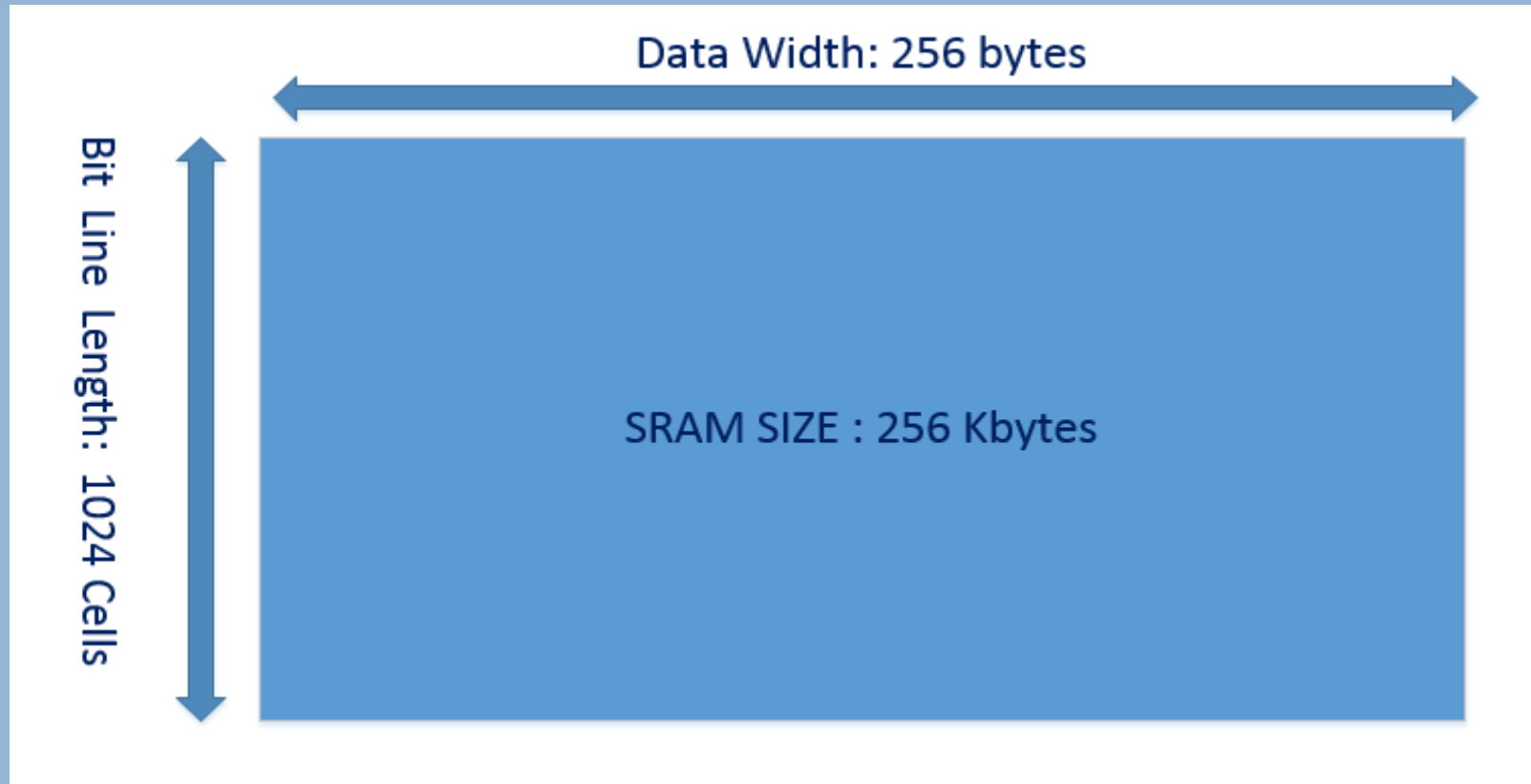


Fig. 1

# L1 Cache New Partition

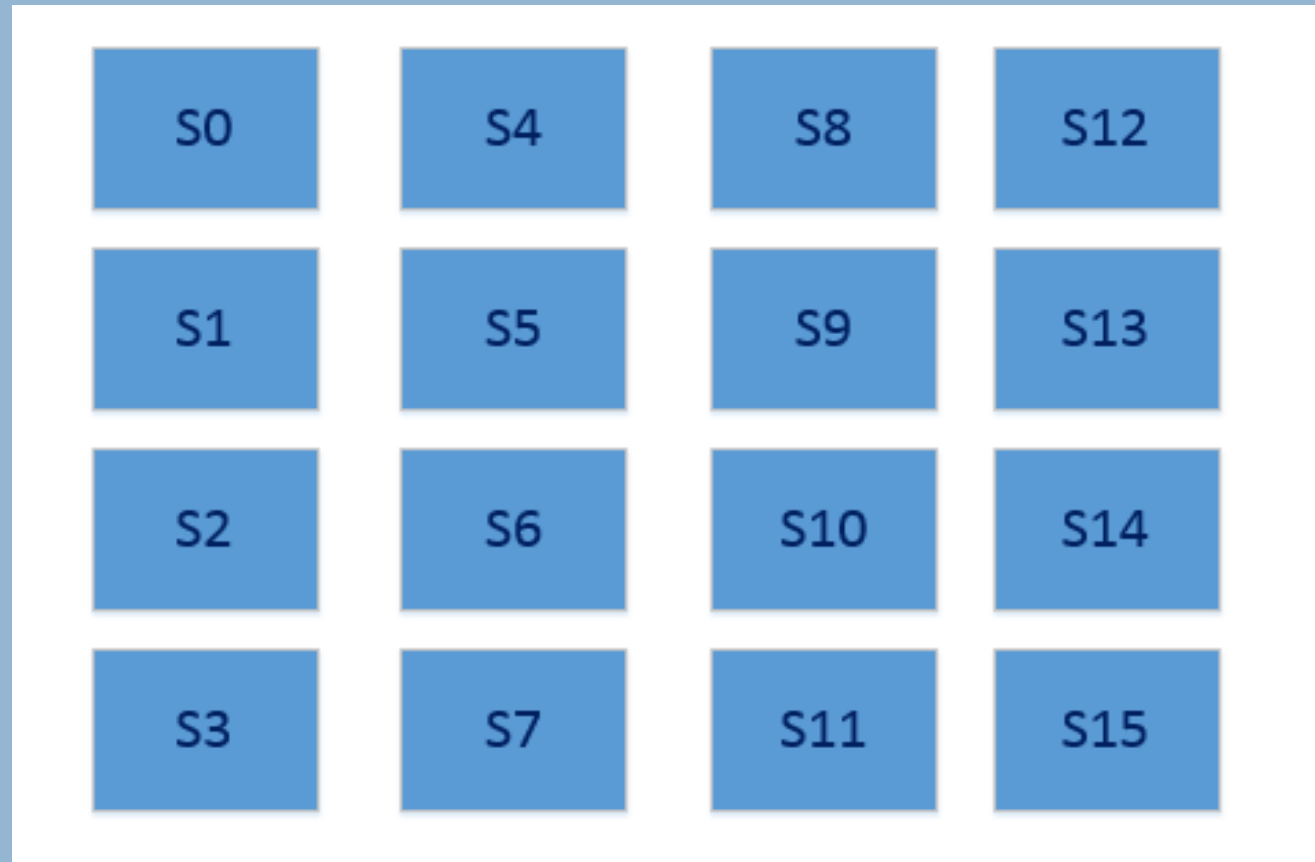


Fig. 2

# Small part of L1 Cache

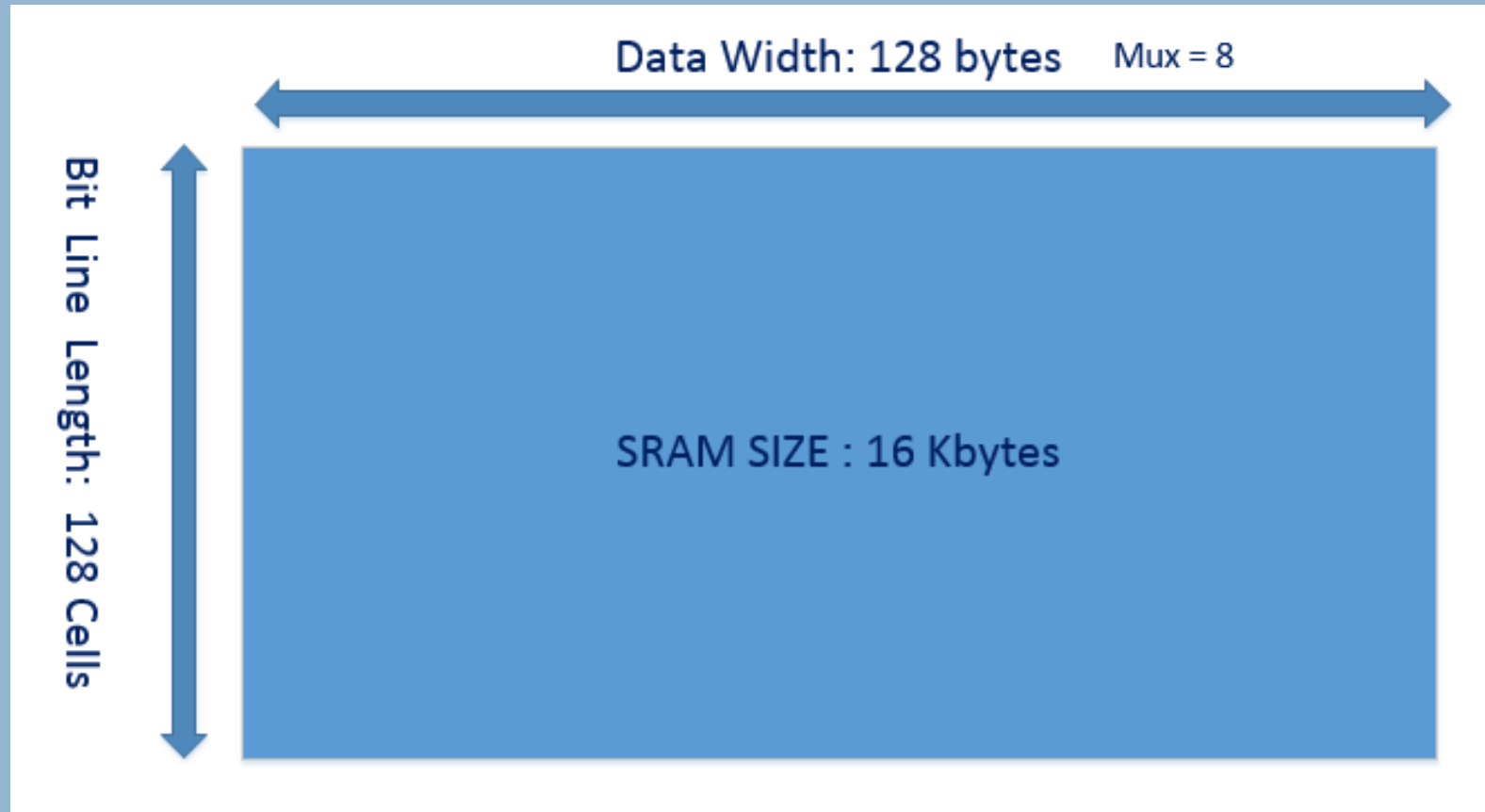


Fig. 3

# The new invention to Small part of L1 Cache

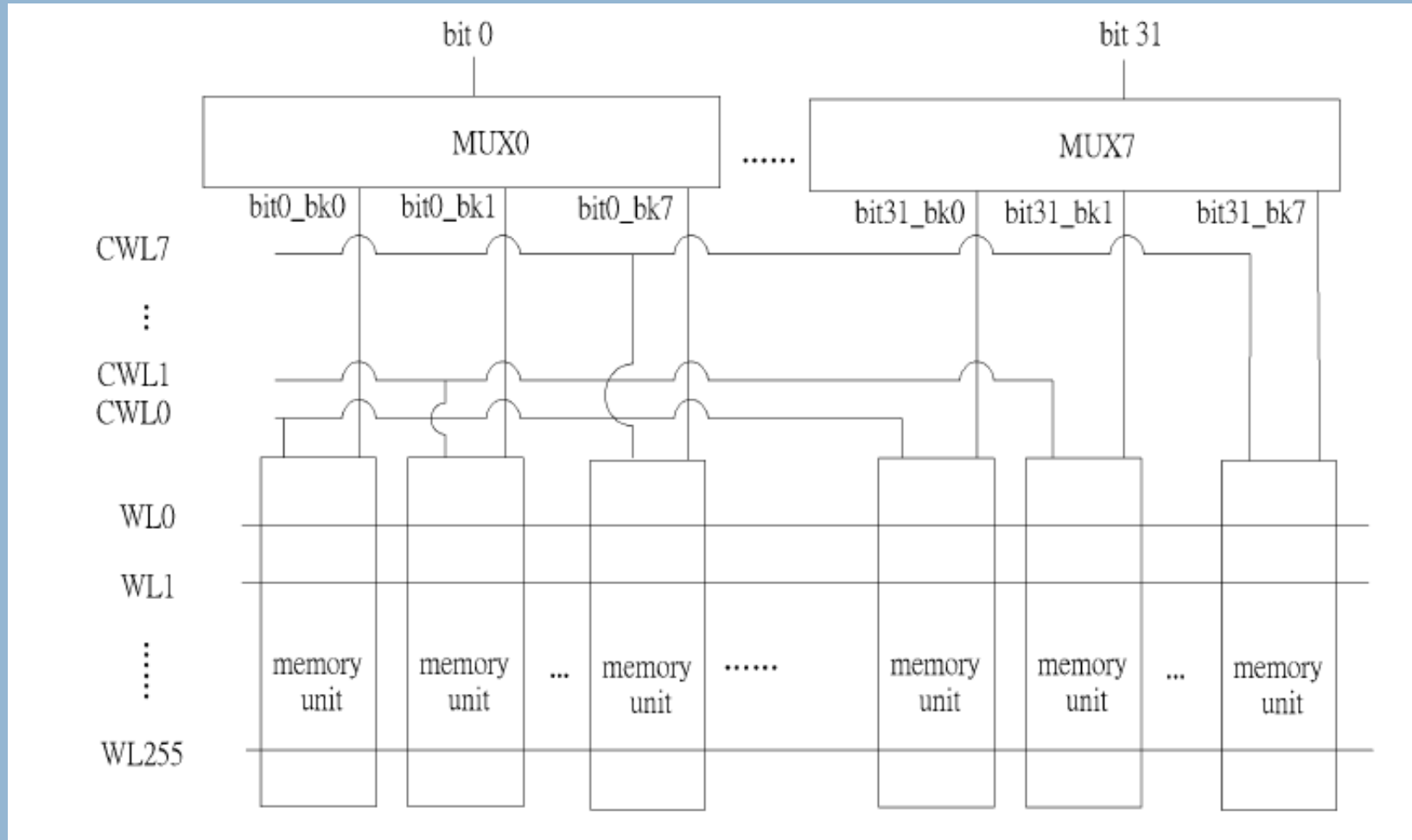


Fig. 4

Note: WL255 change to WL127; bit 31 change to bit 1027 for current size

# Bit line in Each Byte

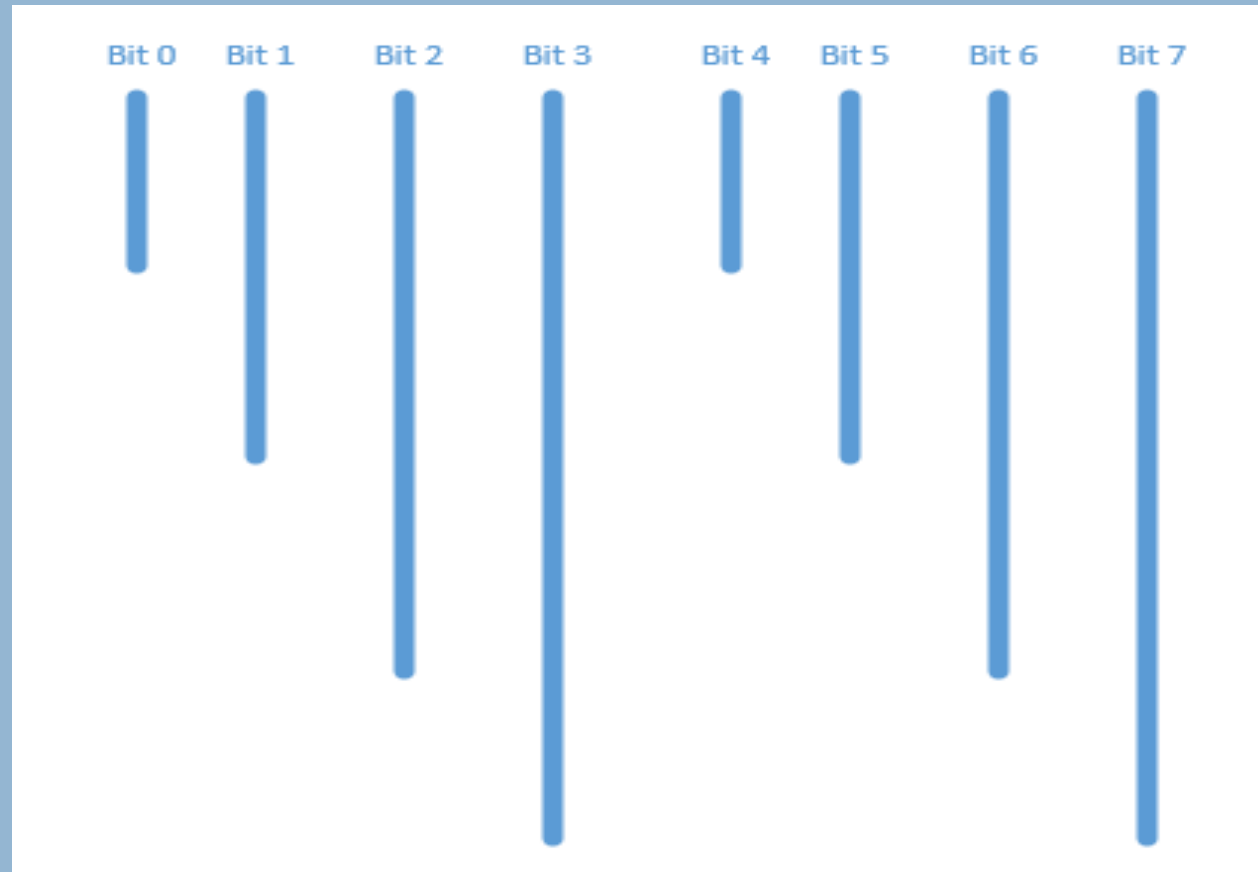


Fig. 5



# Bit line Arrangement

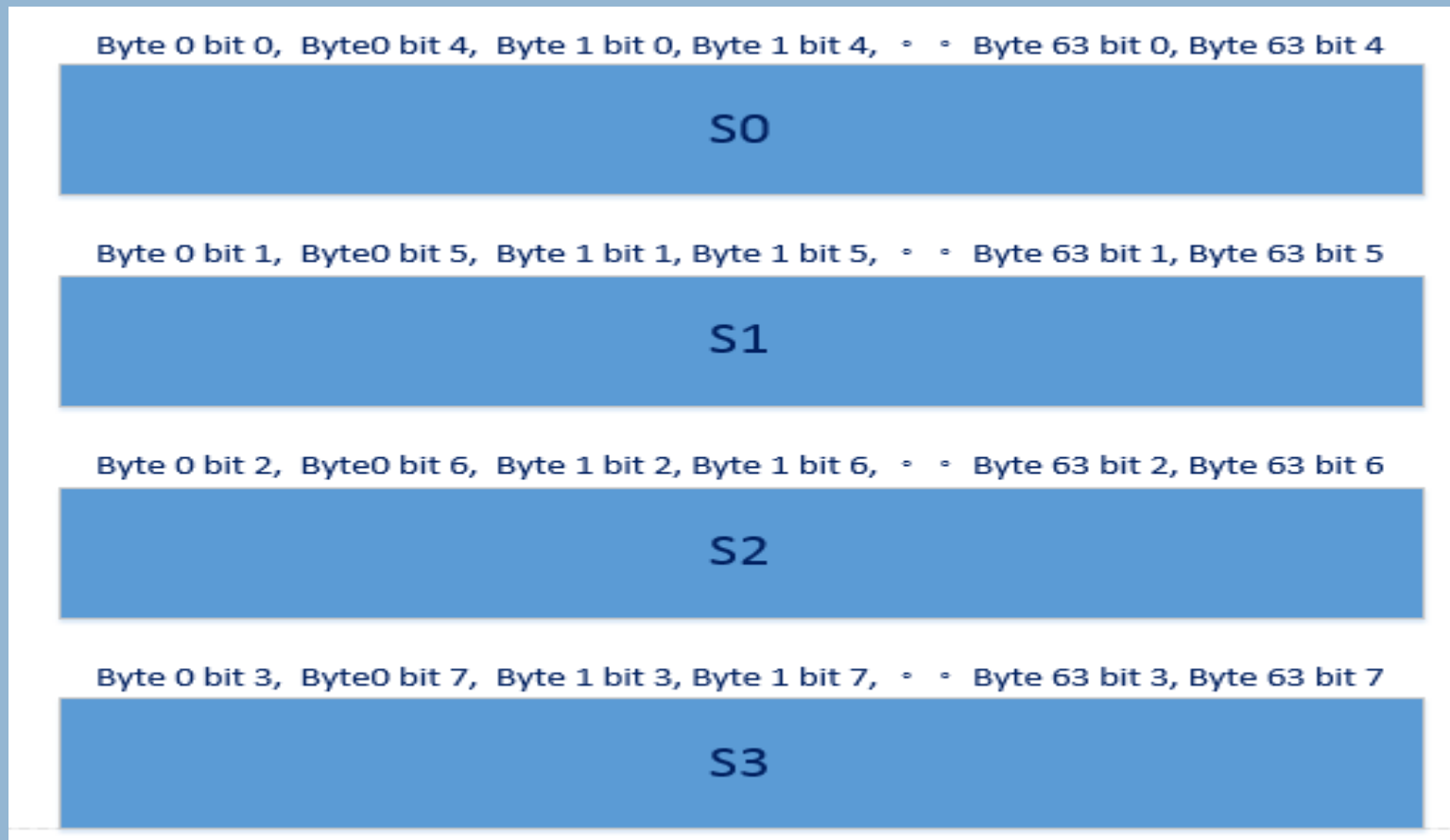


Fig. 6

# Principle I

- For saving bit line power, the original 1024 cells bit line length is divided by 8. Thus each bit line length of each small part is 1/8 of original one.
- Reference Fig. 2 and Fig. 5, the bit line power consumption is  $(1/8 + 2/8 + 3/8 + 4/8)/4 = 31.25\%$  of Fig. 1.
- Due to each bit line of each small part (S0, S1, .... S15) is 1/8 of Fig. 1, so the power and size of sense amplifier can smaller, also the row address decoder smaller from (1024 select 1) to (128 select 1), thus power and size of row address reduces much.

# Principle II

- Reference Fig. 2, ,Fig.5, Fig. 6, bit line output is in the same way(just at S0, S4, S8, S12 TOP), so there does not need more routing power.
- This new partition must use new invention (Fig. 4) to reduce un-selected bit line power consumption(called cell array power here), otherwise new partition can not reduce much power consumption.
- Now bit line power is only 31.25% of original one, and sense amplifier, row decoder power is much smaller, so the new invention power is less 30% of original one.

# Original L2 Cache and New Partition

- Even L2 cache is very large like 40 Mbytes, but L2 cache consists of many banks, and each bank is independent. So only need to focus on L2 bank. The original GPU/AI L2 cache bank is shown in Fig. 7.
- The new invention is shown in Fig. 8. The SRAM is divided to 32 small parts. The size of each small part is shown in Fig. 9. And Fig. 4 is the new invention applied to small part. Bit lines in each byte is shown in Fig. 5.

# Original L2 Cache

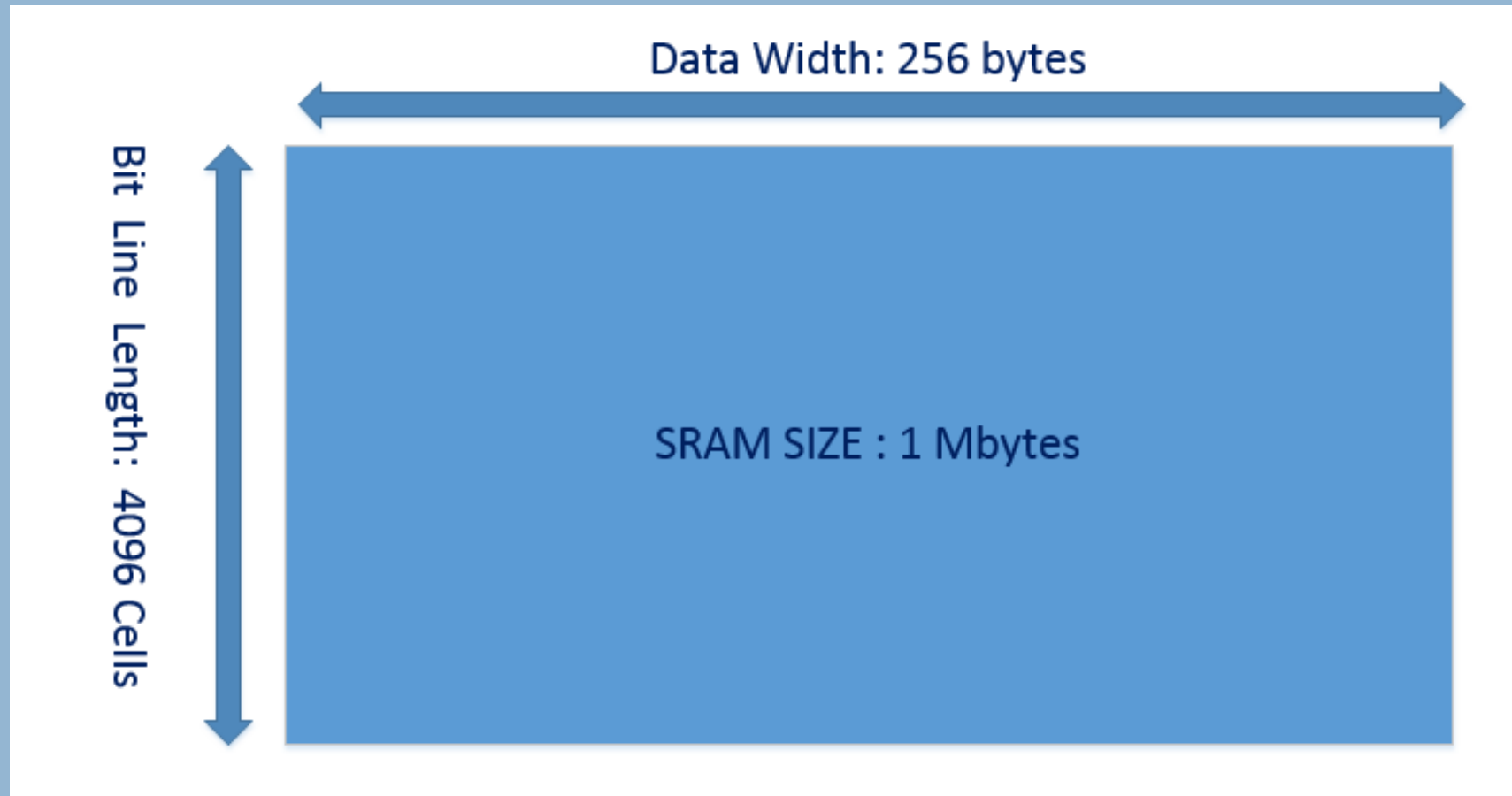
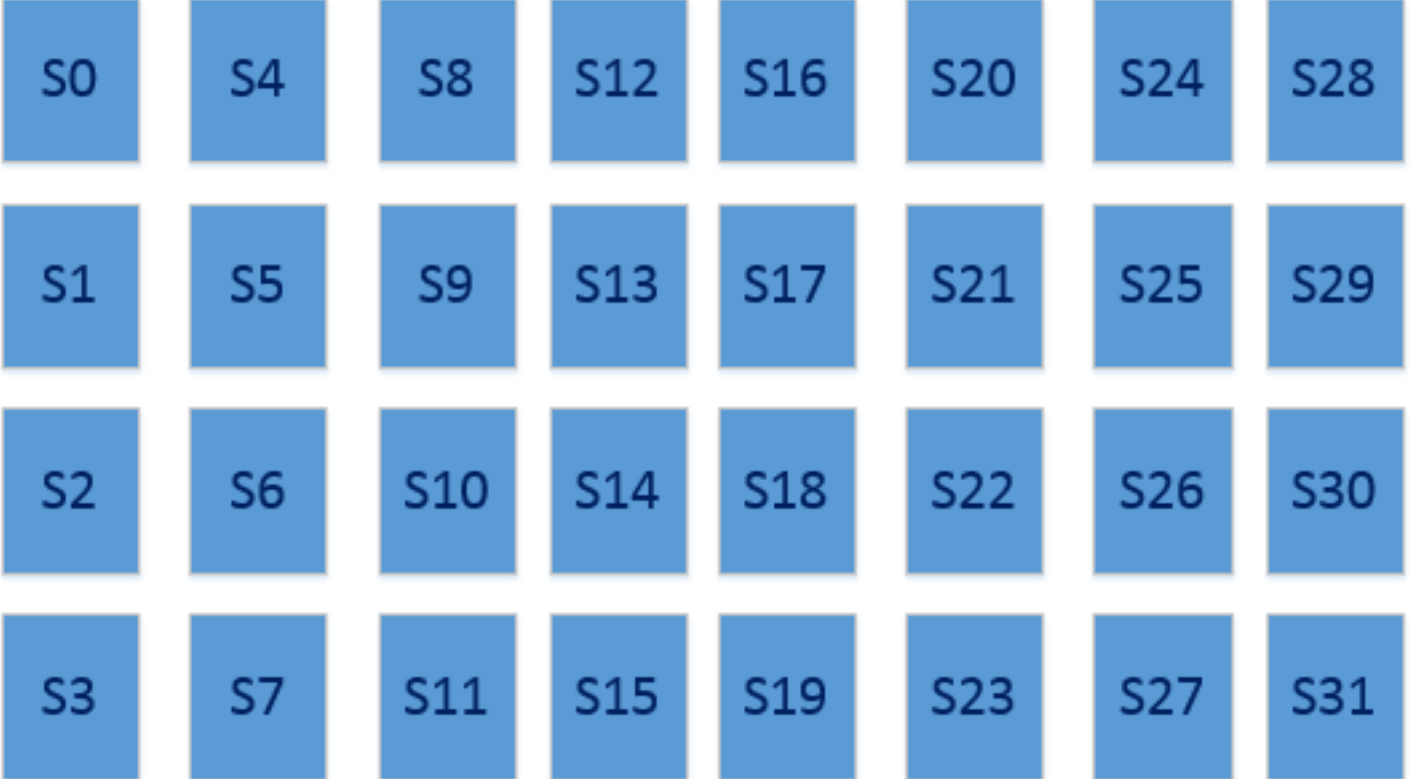


Fig. 7

# L2 Cache New Partition



S0	S4	S8	S12	S16	S20	S24	S28
S1	S5	S9	S13	S17	S21	S25	S29
S2	S6	S10	S14	S18	S22	S26	S30
S3	S7	S11	S15	S19	S23	S27	S31

Fig. 8

# Small part of L2 Cache

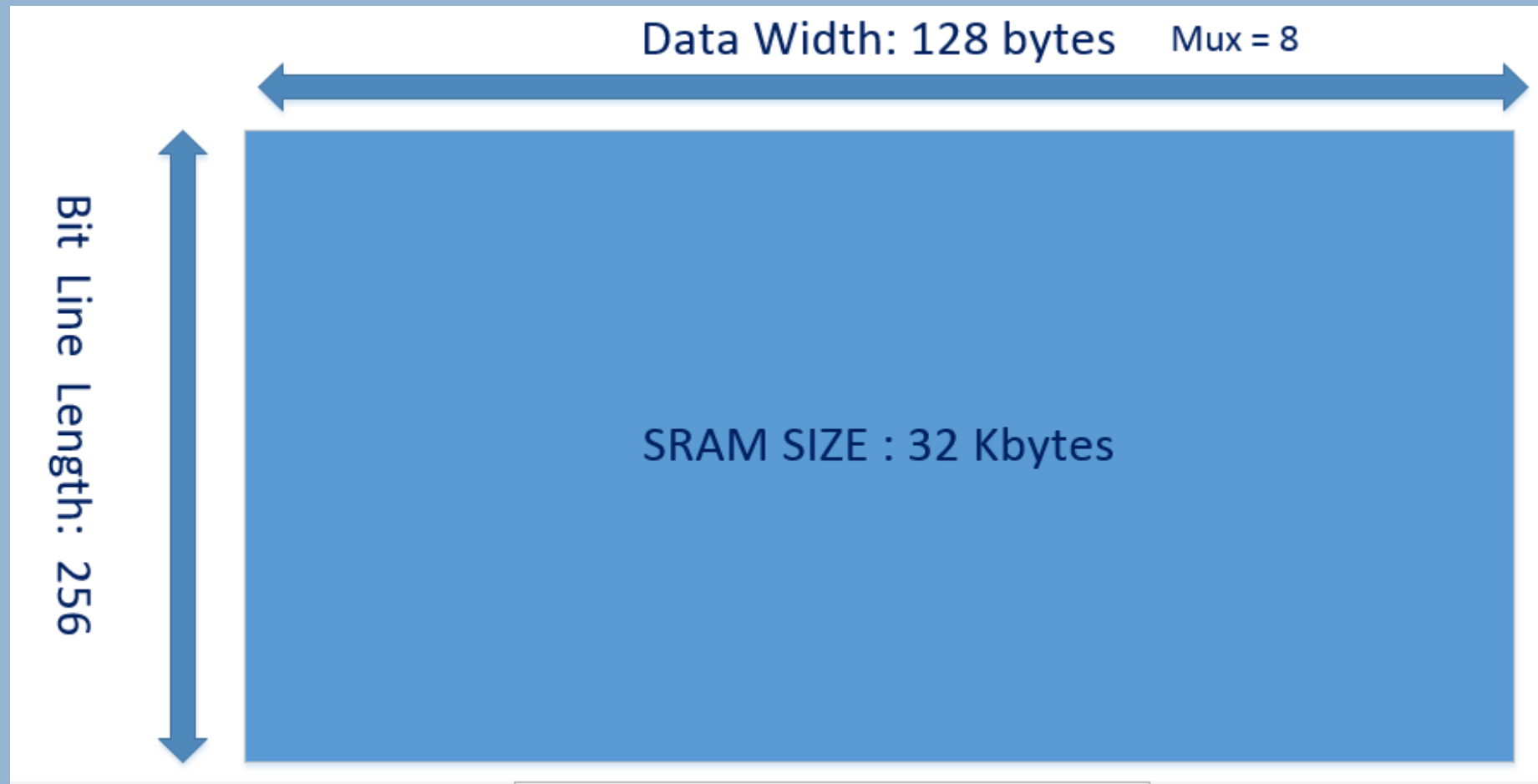


Fig. 9

# Power Consumption of New L2 Cache

- For saving bit line power, the original 4096 cells bit line length is divided by 16. Thus each bit line length of each small part is 1/16 of original one.
- Reference Fig. 8 and Fig. 5, the bit line power consumption is  $(1/16 + 2/16 + 3/16 + 4/16)/4 = 15.625\%$  of Fig. 7.
- So overall power consumption of New L2 cache is less than 15% of original one.



# Notice

- The GPU/AI L1/L2 cache will more and more large in the future product. For L1 cache, it will increase from 256 Kbytes to 1 Mbytes in 3 years, and L2 cache bank will increase from 1 Mbytes to 4 Mbytes.
- With the new invention, this increment will not hurt power efficiency, because the power consumption of L1/L2 cache will smaller than current technology.

Thank You !