Optimized low-precision multiply-and-accumulate for neural computing in pixel sensor applications

Hui Wang, Pengcheng Ai, Dong Wang, Deli Xu, Ni Fang, Fan Shen

1 *Topmetal-M* chip

$Topmetal-M$ (Fig.1) is a large-area pixel array detector chip based on 130nm CMOS process. It has multiple functions of detecting position, energy and time. The analog signal of the pixel detector is digitized by 3-bit precision ADCs in the peripheral circuit, and later the data can be used in subsequent neural computation.

2 Independently designed convolutional neural network chip

We have designed a general-purpose convolutional neural network ASIC (Fig.2), named $PulseDL$, which is currently used in pulse timing tasks for high-energy physics. The newly designed low-precision multiply-and-accumulate is proposed to adapt the chip to high-speed online applications.

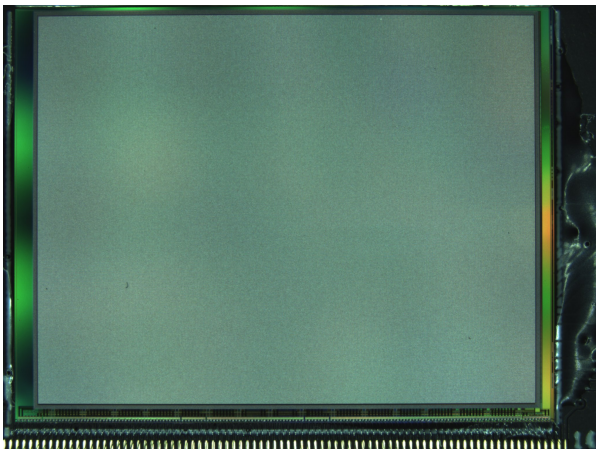

Fig. 1. Photo of $Topmetal-M$.

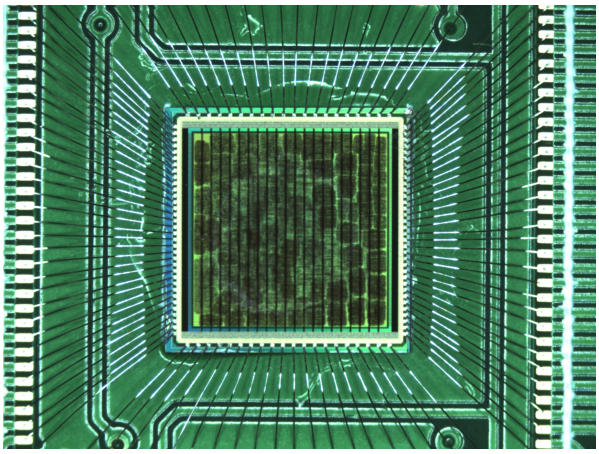

Fig. 2. Photo of $PulseDL$

3 Circuit structure

For low-precision multiply-and-accumulate, the realization of multiplication depends on addition, so we mainly discuss the implementation of the adder. Two major aspects are taken into consideration:

3.1 The number of transistors in the basic full adder

As shown in Fig.3 and Fig.4, full adder can be constructed by twenty-four transistors and twenty-eight transistors, respectively. The performance and power consumption are almost the same. Compared to the thirty-six transistors full adder that has not been optimized to reduce the number of transistors, the performance of the former two is improved by more than 70%, and the power consumption is reduced by 15% (15% lower in power).

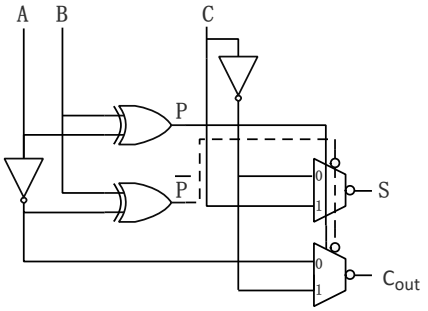

Fig. 3. Twenty-four transistors full adder

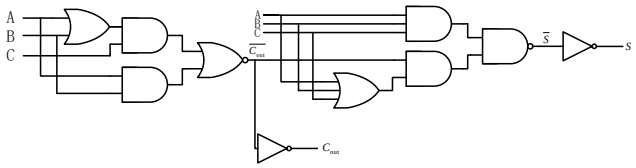

Fig. 4. Twenty-eight transistors full adder

3.2 Hybrid model of classic addition structures for low-precision data

Here, as shown in Fig.5, we have depicted the gate structure of a 3-bit adder, and our design expectation is to find the best balance point of power, performance and area.

In this circuit structure, we use reserved carry addition located at the first and the third stages, respectively, and a carry look-ahead adder at the second stage. The gate circuit structure is designed according to the following formula:

Generate(G) : $G_i = A_i B_i$

Propagate(P) : $P_i = A_i \oplus B_i$

First level: $S_0 = G_0 + P_0 C_{-1}$ $C_0 = G_0 + P_0 C_{-1}$

Second level: $S_1 = G_1 + P_1 C_0$ $C_1 = G_1 + P_1 G_0 + P_1 P_0 C_{-1}$

Third level: $S_2 = G_2 + P_2 C_1$ $C_2 = G_2 + P_2 C_1$

After analysis, this structure has a higher performance than the reserved carry adder and can pass through one less gate circuit. At the same time, it has fewer gate circuits and a better area than the look-ahead adder.

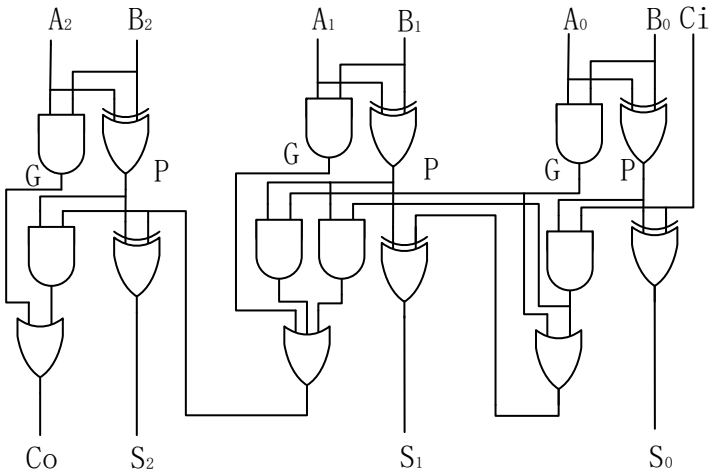

Fig. 5. Proposed 3-bit adder

4 Conclusion

In the entire multiply-and-accumulate, in order to enhance the performance of the multiplication part, we have designed the 3-bit encoding multiplication and the 3-bit look-up table multiplication. Then in the partial product processing, a new 3-bit adder is designed. These improvements have improved computing performance and reduced power consumption and area.