Нейропроцессор NM6403

NeuroMatrix® Engine

Версия 1.0

АКЦИОНЕРНОЕ ОБЩЕСТВО
НАУЧНО-ТЕХНИЧЕСКИЙ ЦЕНТР

This page was intentionally left blank.

Архитектура NeuroMatrix предоставляет уникальную гибкость в выборе требуемого уровня производительности и точности для процедур умножения с накоплением (MAC). Исходя из требований приложения, можно выбрать необходимую длину операндов и результатов упакованных в 64-бит слова данных. Количество MAC будет зависеть от длины и количества операндов и результатов. Наивысшая производительность достигается в случае использования 1-бит операндов. В этом случае, при тактовой частоте 50МГц производительность составляет 51200 MMAC (миллионов логический умножений с накоплением). Для увеличения точности вычислений, можно использовать операнды длиной до 32-бит. Длина операндов может быть любой, даже не кратной степени двойки. В предельном случае, 32-бит операнды и 64-бит результат, производительность составит 50 MMAC (миллионов арифметических умножений с накоплением). Такой подход позволяет делать выбор между точностью вычислений и их производительностью. Пример конфигурации NeuroMatrix Engine, для случая умножения матрицы байтов на вектор байтов показан на рис. 1.

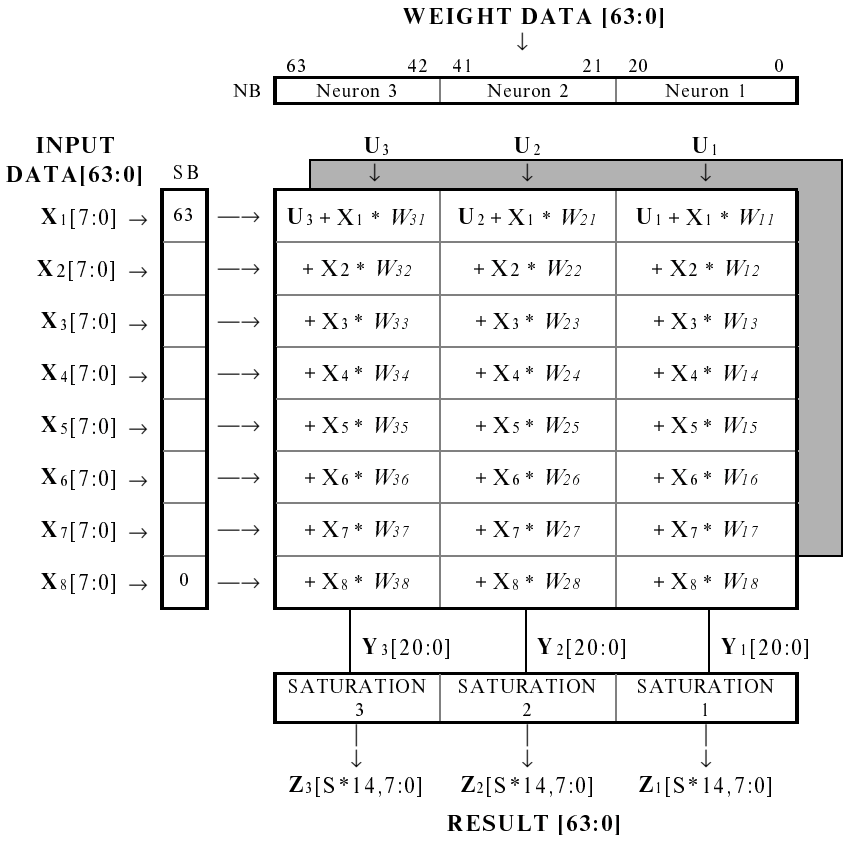
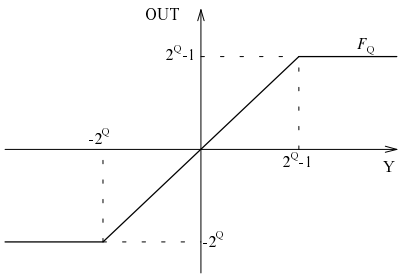


рис. 1

Ядром архитектуры является регулярная структура похожая на матричный умножитель. Матрица содержит 64x64 ячеек, каждая ячейка содержит элемент памяти (flip-flop) м несколько логических элементов. Матрица может быть разделена на несколько подматриц двумя 64-бит программируемыми регистрами: SB и NB. Эти регистры определяют границы синапсов и нейронов соответственно. Например, для 8-бит синапсов (Xi) и 8-бит весов (Wi) количество подматриц (макроячеек) составляет 24. Каждая макроячейка производит операцию умножения элементов вектора входных данных Xi на предварительно загруженные коэффициенты Wi и накапливает результат из макроячеек расположенных выше и входа Ui. Таким образом, каждый столбец вычисляет 21-бит выход нейрона имеющий 8 байтных синапсов и 8 байтных весовых коэффициентов. В нашем случае, имеется три таких нейрона. Значения выходов нейронов, вычисляются параллельно за один процессорный такт. При этом производится 24 операции MAC. При значении тактовой частоты 50МГц, производительность составляет 1200 MMAC .

На загрузку весовых коэффициентов в матрицу, требуется 32 такта. Для снижения накладных расходов, связанных с перезагрузкой матрицы весов, применяется “теневая” матрица. Новые данные загружаются в “теневую” матрицу на фоне вычислений и могут быть переданы в рабочую матрицу за один процессорный такт.

Функция насыщения (Saturation Function) используются для снижения разрядности результатов и защиты от арифметического переполнения.



Saturation Function

В примере приведенном на рис. 1, функция насыщения снижает количество значащих бит с 21-бит до 8-бит. Ширина входов функции насыщения эквивалентна ширине колонки (выхода нейрона), ширина выхода функции должна быть эквивалентна входу нейрона. На первом проходе, функции насыщения снижают количество значащих бит, на втором проходе, рабочая матрица упаковывает 8-бит выходы в 64-бит слова данных. Все параметры функций насыщения программируются.

Конфигурация матрицы может быть изменена динамически в течение вычислений. Вычисления могут быть начаты с максимальной точностью и минимальной производительностью, но при определенных условиях можно достичь пиковой производительность путем снижения точности.

Для вычисления производительности используются следующие выражения:

$$MCPS = \left\lfloor \frac{64}{N_X} \right\rfloor * \left\lfloor \frac{64}{N_X + N_W + \left\lceil \log_2 \frac{64}{N_X} \right\rceil} \right\rfloor * F \ ,$$

где: MCPS - миллион соединений в секунду (эквивалентно MMAC)

64 - ширина слова данных;

N_X - ширина синапсов;

N_W - ширина весов;

F = 50 МГц - тактовая частота.

В случае $N_X \neq 1$ и $N_W = 1$, выражение приобретает вид:

$$MCPS = \left\lfloor \frac{64}{N_X} \right\rfloor * \left\lfloor \frac{64}{N_X + \left\lceil \log_2 \frac{64}{N_X} \right\rceil} \right\rfloor * F \ .$$

В случае $N_X = 1$ и $N_W \neq 1$, выражение приобретает вид:

$$MCPS = 32 * \left\lfloor \frac{64}{N_Y + 5} \right\rfloor * F \ .$$

В случае $N_X = N_W = 1$, выражение приобретает вид:

$$MCPS = 1024 * F$$

И наконец, в случае $N_X = N_W = 32$, выражение приобретает вид:

$$MCPS = F$$

Некоторые результаты вычисления соотношения производительность /точность приведены на диаграмме показанной на рис. 2.

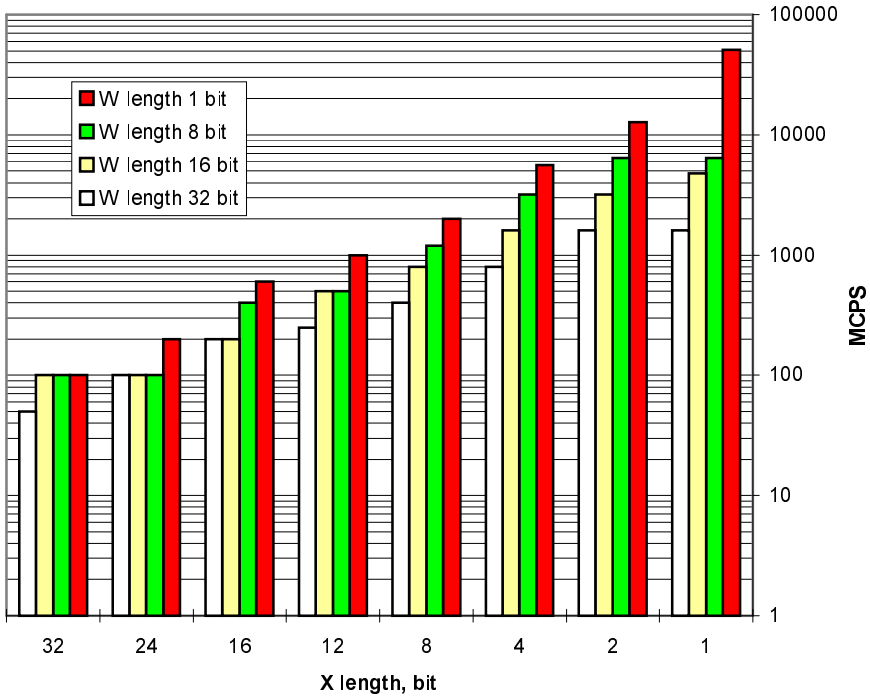


рис. 2

Как видно из приведенной диаграммы, программист может выбрать желаемую точку в диапазоне: 50 - 51200 MMAC(MCPS).

This page was intentionally left blank.

Научно-технический центр Модуль
АЯ 166, Москва, 125190, Россия
Тел: +7 (095) 152-9335
Факс: +7 (095) 152-4661
E-Mail: postmast@module.vympel.msk.ru
WWW: http://www.module.vympel.msk.ru

Напечатано в России. Дата издания: 23 февраля 1999