به نام خدا



# فاز سوم پروژه

طراحی سیستم های دیجیتال

استاد فلاحتى

# اعضای تیم:

سهیل نظری مندجین علیرضا حقی مهدی سعادت بخت

نیمسال دوم سال تحصیلی ۱۴۰۰-۱۴۰۱ دانشگاه صنعتی شریف دانشکده مهندسی کامپیوتر

# فهرست مطالب

3	توضيحات پروژه
3	توضیحات پروژه چکیده
4	مقدمه
4	کاربردهای ضرب دو ماتریس
4	روشهای ضرب دو عدد
4	تقسيم و غلبه
	درخت والاس
6	ضربکننده بران (آرایهای)
7	FSM and ASM
9	الگوريتم ضرب ماتريس
10	مرحله اول و دوم (پارتیشنبندی ماتریس به ساببلاکهای سایز $(p * np * np )$
10	مرحله سوم (ضرب ساببلاکها در هم و سپس جمع آنها با هم)
11	مرحله چهارم (شیفت دادن ماتریس a به راست و ماتریس b به پایین)
	ضرب ماتریس
12	سنتز
12	RTL Schematic
	Synthesis schematic
13	نحوه برخور د با اعداد اعشاری و تست برنامه.

## توضيحات يروژه

یک واحد پردازشی ضرب دو ماتریس را با استفاده از الگوریتم تقسیم و حل پیاده سازی کردهایم. در این طراحی مولفههای ماتریس مطابق با استاندارد IEEE754 دریافت میشوند. این طراحی با استفاده از تقسیم و حل به گونهای انجام شده که انجام عملیات ضرب ماتریس به صورت موازی و بهینه انجام شود.

#### چکیده

با توجه به محاسبات سنگین محاسباتی ماتریسی در حوزههای مختلف، پردازش موازی ضرب ماتریسها در راستای کمینه کردن میزان انرژی و مساحت مورد استفاده اهمیت ویژهای پیدا کرده است.

در این پروژه، بنابر دغدغهی الزام ضرب موازی ماتریسها، به طراحی یک مدل تقریبا بهینه با استفاده از زبان برنامهنویسی Verilog بر اساس مساحت و تعداد ماژولها رسیده ایم. در مدل ما از  $O(\frac{n^3}{p})$  واحدهای محاسباتی استفاده شده است که قادر است در O(p) کلاک محاسبات موازی دو ماتریس را انجام دهد که در این جا p بیانگر مجذور تعداد واحدهای پردازش موازی است. کدهای ما کاملا قابل سنتز است و این امر به وسیلهٔ نرمافزار سنتز Vivado صورت گرفته است. همچنین صحت کد و کارایی آن توسط آزمایش کننده ای توسط دو کد پایتون که یکی عدد رندوم تولید کرده و دیگری پاسخ آن را میابد صورت گرفته است. نکته می حائز اهمیت دیگر مدل ما، پارامتریک بودن است که می توان به وسیلهٔ آن برای هر ابعاد دلخواه و هر تعداد دلخواه پردازنده آن را در نظر گرفت که در آن تعداد پردازنده مربع کامل است و ابعاد بر مجذور پردازنده ها بخش پذیر است.

#### مقدمه

#### کاربردهای ضرب دو ماتریس

ضرب دو ماتریس در علوم کامپیوتر کاربردهای گستردهای دارد و از اساسی ترین دلایلی که سعی داریم تا این عمل را در پایین ترین سطح پر دازشی، یعنی سطح تر انزیستور و چیپ انجام دهیم و از انجام آن به صورت نرمافزاری خودداری کنیم، کاربردهای گسترده آن است. از گسترده ترین استفادههای ضرب دو ماتریس می توان در گرافیک یاد کرد. هر تصویر دیجیتالی را می توان با یک ماتریس مدل کرد (همانطور که در حال حاضر نیز این امر صورت می پذیرد). هر خانه از این ماتریس را می توان حاوی اطلاعات یکی از پیکسلهای این عکس در نظر گرفت که عدد آنها نشان دهنده ترکیب رنگی آن پیکسل است. حال برای مثال، برای دیکود کردن یک ویدیو دیجیتالی، نیاز به ضرب ماتریسها داریم. بسیار دیده ایم که برای مثال یک ویدیو را می توان از فرمت 254 به 255× دیکود کردن یک کرد و این امر با توجه به بالا بودن نرخ فریمها در یک ثانیه برای ویدیوهای جدید، اگر میخواست در سطح نرمافزار انجام می شد، زمان بسیار زیادی را از کاربر میگرفت. پژوهشگران MIT یکی از اولین چیپهای مخصوص (ASIC) و بهینه برای کد کردن کامپیوتر با استفاده از اصطلاح پردازش دیجیتال سیگنال بیان می شود. از دیگر کاربردهای ضرب ماتریسها خلق کرده اند. گرافیک در تصویربرداری دیجیتال، پردازش سیگنالها و چندرسانه ای اشاره کرد. همچنین در کاربردهای جدیدتر برای شتابنده هی سخت افزاری هوش مصنوعی، ضرب ماتریس در مقیاس بالا استفاده می شود. یک تصویربرداری دیجیتال، پرداز ش سیگنالها و چندرسانه ای اشاره کرد. همچنین در کاربردهای جدیدتر برای شتابنده های سخت افزاری هوش مصنوعی، ضرب ماتریس در مقیاس بالا استفاده می شود.

#### روشهای ضرب دو عدد

برای ضرب دو عدد، روشهای متعددی وجود دارد و انتخاب هر کدام از آنها، مسئله انتخاب بین سرعت و استفاده از سختافزار کمتر است. به طوری که برخی از روشها با وجود سرعت بالا در محاسبه، نیازمند تعداد بالایی منابع سختافزاری هستند، اما برخی دیگر میتوانند این کار را در زمانی طولانی تر اما با منابع سختافزاری محدود تر انجام دهند. ۳ روش معروف برای این کار که به بررسی آنها خواهیم پرداخت عبارتند از:

- تقسیم و غلبه (Conquer & Divide)
  - درخت والاس (Tree Wallace)
- ضرب کننده بران (Multiplier Braun)

#### تقسيم و غلبه

این روش بیشتر در FPGAها به کار میرود که همین امر باعث سرعت بیشتر میشود. به دلیل موازیسازیهای که در این روش انجام میشود، مساحت به کار رفته در آن و همچنین قطعات مورد استفاده میتوانند بیشتر شوند. این روش به روش کاراتسوبا نیز معروف است. طرز کار این الگوریتم به این صورت است که ابتدا هر کدام از دو عدد را به نیم تقسیم کرده. بر اساس تقسیمبندی ای که کردیم، عبارت نهایی پس از ضرب باید به شکل زیر باشد:

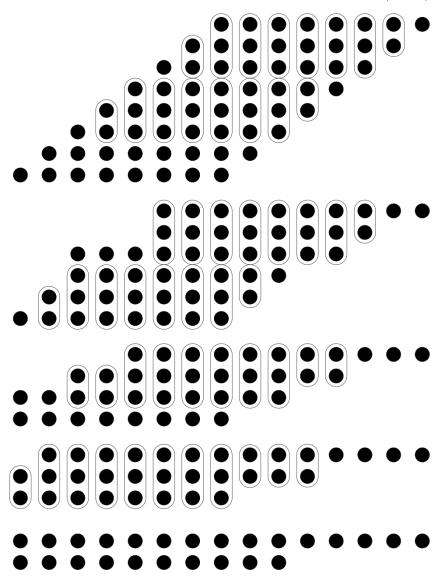
Ah. Bh. 
$$10^n$$
 + AlBl + AhBl $10^{\frac{n}{2}}$  + AlBh $10^{\frac{n}{2}}$ 

سپس ۳ عبارت زیر را محاسبه میکنیم:

بعد از محاسبه این  $^{\circ}$  عبارت، میتوانیم به سادگی عبارت نهایی را محاسبه کنیم. پیچیدگی زمانی نهایی این محاسبات برای رسیدن به جواب نهایی، از مرتبه زمانی  $O(n^{\log(3,2)})$  خواهد بود.

#### درخت والاس

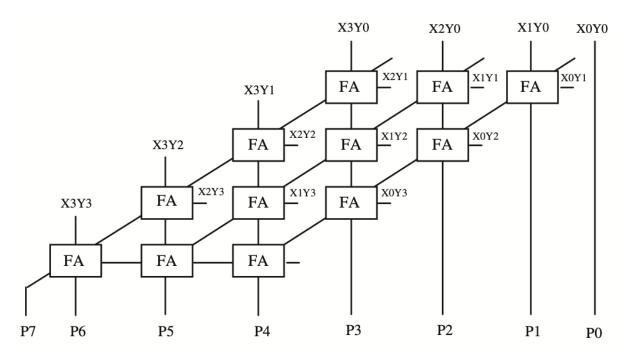
درخت والاس در هر سطح از ضرب، تعداد بیتهای کمتری دارد ولی تاخیر به صورت موازی تقسیم نمی شود. با توجه به ارتباطات میانی بسیاری که درخت والاس دالاس دارد، مساحت این روش زیاد می شود. صورت کلی انجام این روش برمی گردد به جمع ریز-ضربهایی که انجام می دهیم. شکل کلی آن در زیر قابل مشاهده است:



تصویر بالا مربوط به یک درخت و الاس ۸\*۸ است که ۴ لایه از آن کاهش یافته است. هر کدام از نقطه ها نمایانگر یک بیت باارزش و زن یکسان هستند. این عمل کاهشی با استفاده از adder-half ۱۴ و adder-full ۳۸ اوزن یکسان هستند. این عمل کاهشی با استفاده از adder-half ۱۴ و محانطور که در ابتدای این بخش نیز بیان شد، کاهش کردن دو نقطه نشان داده شده و adder-full ها با پکیج کردن سه نقطه. همان طور که در ابتدای این بخش نیز بیان شد، کاهش زمانی انجام یک عملیات (که در اینجا با کاهش لایه ها قابل مشاهده است)، به قیمت افز ایش استفاده از منابع سخت افز اری است.

## ضرب کننده بران (آرایهای)

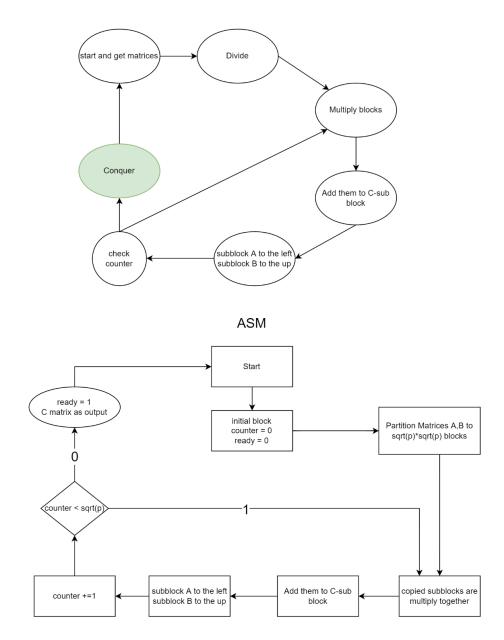
این روش برخلاف روش قبل، توزیع تاخیر یکسانی برای تمام سطوح دارد. این روش در اصل همان ضرب آرایهای است که از دوران مدرسه آن را آموختهایم. شکل آن به صورت زیر است:



این روش از بدترین پیچیدگی زمانی با مرتبه زمانی  $O(n^3)$  برخوردار است. اما مزیت استفاده از این روش، کاهش پیچیدگی پیادهسازی و استفاده از منابع سختافزاری خواهد بود.

#### **FSM** and ASM

در این بخش ما شمایی کلی از حالت و وضعیتهای محاسباتی ممکن برای ماژول ضربکننده نشان میدهیم که از ۷ حالت تشکیل شده است:



با استفاده از طراحی FSM انجام شده یک ماژول کنترلر تعریف کردهایم که به وسیله آن استیتهای مختلف ماشین ما تغییر کرده و سه بیت کنترلی برای خواندن، جمع کردن و شیفت دادن در اختیار داریم که میان استیتها مقادیرشان را تنظیم میکنیم. نکتهٔ دیگر این است که ماشین کنترلی مد نظر از نوع Moore است. بنابراین به ورودی برای حرکت بین حالات حساس نیست.

#### تکه کد ریست و استارت در استیت صفر:

```
always @(posedge clk, posedge reset)
begin
    if (state == 0)
    begin
        nx_state = 1;
        enable_read = 1;
    end
    if(reset == 1)
        state <= 0;
    else
        state <= nx_state;
end</pre>
```

#### تکه کد انتقال میان استیتها و ست کردن بیتهای کنترلی:

```
always @(state)
always e
begin
if(!reset)
begin
if(enable)
case(st
                         case(state)
0:
                                                 nx_state = 1;
enable_read = 1;
                                          begin
                                                 enable_read = 1;
nx_state = 2;
                                         enable_read = 0;
enable_sum = 1;
nx_state = 3;
end
                                                  enable_sum = 0;
enable_shift = 1;
nx_state = 4;
                                                  enable_shift = 0;
if (shifted < sqrt_p - 1)
nx_state = 2;
                                                  nx_state = 5;
shifted = shifted + 1;
                                                 out_ready = 1;
nx_state = 6;
                                         out_ready = 0;
nx_state = 0;
end
                                  default:
begin
                                               out_ready = 0;
enable_shift = 0;
enable_sum = 0;
enable_read = 0;
nx_state = 0;
                  else
```

# الكوريتم ضرب ماتريس

الگوریتم کلی بدین شکل است که هر دو ماتریس n \* n را به  $p^2$  ماتریس با ابعاد  $\frac{n}{p} * \frac{n}{p}$  تقسیم میکنیم. سپس در طی p مرحله با استفاده از شیفت و ضرب اجزای متناظر ماتریسهای به دست آمده و در نهایت جمع آن ها جواب نهایی حاصل می شود. در عکس پایین نیز شبه کدی برای الگوریتم ارائه شده است  $p^2$  را تعداد پردازنده ها است).

عملیاتهای ضرب دو ماتریس به شکل تقسیم و حل به ۵ مرحله تقسیم میشود که به شکل زیر است.

# Implementation

- Consider two square matrices A and B of size n that have to be multiplied:
  - 1. Partition these matrices in square blocks p, where p is the number of processes available.
  - Create a matrix of processes of size p<sup>1/2</sup> x p<sup>1/2</sup> so that each process can maintain a block of A matrix and a block of B matrix.
  - 3. Each block is sent to each process, and the copied sub blocks are multiplied together and the results added to the partial results in the C sub-blocks.
  - 4. The A sub-blocks are rolled one step to the left and the B sub-blocks are rolled one step upward.
  - 5. Repeat steps 3 y 4 sqrt(p) times.

ما در یک ماژول به اسم array-divider این مراحل را با استفاده از جنریتورها پیادهسازی کردهایم. تکهکد مراحل مختلف به ترتیب آمدهاند. در این بخش بهواسطهٔ جنریتورها توانستیم برنامهٔ کاملا کارا و پارامتریک که هیچ وابستگی به متغیر و عدد ثابتی ندارد را پیادهسازی کنیم که هر چند منجر به پیچیدگیهای قابل ملاحظهای در پیادهسازی کد شد.

# $(\frac{n}{p} * \frac{n}{p})$ مرحله اول و دوم (پارتیشنبندی ماتریس به ساببلاکهای سایز

```
generate

for (1 = 9; i + sqrt_p; i = i + 1)

for (1 = 9; i + sqrt_p; j = j + 1)

for (2 = 9; i + sqrt_p; j = j + 1)

for (3 = 9; i + sqrt_p; j = j + 1)

for (4 = 9; i + sqrt_p; j = j + 1)

for (4 = 9; i + sqrt_p; j = j + 1)

for (4 = 9; i + sqrt_p; j = j + 1)

for (4 = 9; i + sqrt_p; j = j + 1)

for (5 = 9; i + sqrt_p; j = j + 1)

for (6 = 9; i + sqrt_p; j = j + 1)

for (6 = 9; i + sqrt_p; j = j + 1)

for (6 = 9; i + sqrt_p; j = j + 1)

for (6 = 9; i + sqrt_p; j = j + 1)

for (6 = 9; i + sqrt_p; j = j + 1)

for (7 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (8 = 9; i + sqrt_p; j = j + 1)

for (9 = 9; i + sqrt_p; j = j + 1)

for (9 = 9; i + sqrt_p; j = j + 1)

for (9 = 9; i + sqrt_p; j = j + 1)

for (9 = 9; i + sqrt_p; j = j + 1)

for (9 = 9; i + sqrt_p; j = j + 1)

for (9 = 9; i + sqrt_p; j = j + 1)

for (1 = 9; i + sqrt_p; j = j + 1)

for (1 = 9; i + sqrt_p; j = j + 1)

for (1 = 9; i + sqrt_p; j = j + 1)

for (1 = 9; i + sqrt
```

#### مرحله سوم (ضرب ساببلاکها در هم و سپس جمع آنها با هم)

#### مرحله چهارم (شیفت دادن ماتریس a به راست و ماتریس b به پایین)

مرحله پنجم نیز در تصاویر نشان داده است. همانطور که میبینید داخل یک حلقه به اندازه  $\sqrt{p}$  قرار دارند.

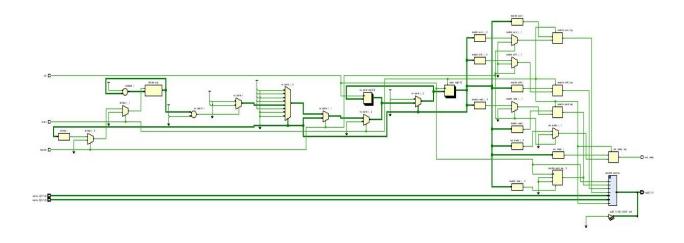
#### ضرب ماتریس

یکی دیگر از بخشهای مورد نیاز برای پیادهسازی ضرب موازی ماتریسها، ضرب دو ماتریس با ابعاد دلخواه است که در آن برای ضرب دو ماتریس n\*n از  $O(n^3)$  واحد ضرب و واحد جمع استفاده شده است که delay که در این عملیات وجود دارد برابر است با n برابر delay یک واحد جمعکننده به علاوهٔ یک واحد ضربکننده. هر چند اگر بهینه تر پیادهسازی کرد، می توان این delay را به  $O(\log(n))$  بار delay جمعکننده و یک delay ضربکننده رساند که با توجه به بزرگ تر شدن n می توان به شکل قابل توجهی سرعت کلاک را کمتر کرد و در غیر این صورت می توان سرعت کلاک را پایین نگه داشت ولی تعداد دور هایی که کلاک در استیت ضرب کردن است بنابر آزمانی به دست آورد.

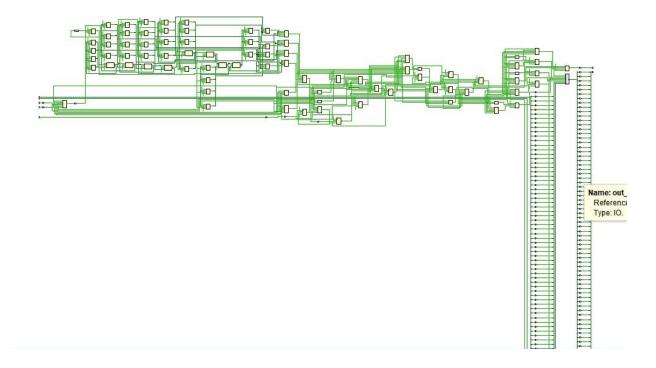
#### سنت

ما با استفاده از ابزار سنتز Vivado طراحی خود در وریلاگ را سنتز کرده و کدهای ما کاملا سنتزپذیر بوده و خروجیهای آن به شکل زیر میباشد.

#### **RTL Schematic**



# Synthesis schematic



ما با استفاده از روش تقسیم و حل و همچنین موازی سازی عملیاتهای پیدا کردن هر یک از درایههای ماتریس خروجی در توان مصرفی نهایی قطعه خود صرفه جویی بسیاری انجام دادهایم. طبق آزمایشها و محاسباتی که بر روی این قضیه انجام شده است، جدولهای زیر برای سه الگوریتم مختلف انجام این کار نشان داده شده است.

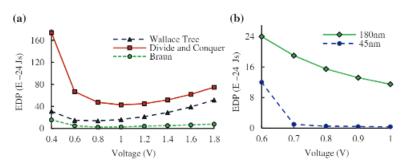


Fig. 4 a EDP of different multipliers at 180 nm technology node. b EDP comparison of Barun multiplier at 180 and 45 nm technology nodes

#### **EDP**

The proposed work uses energy-delay product (EDP), where energy the total energy consumption of cores and delay is the amount of time for executing applications.

نتیجه یکی از تست های امتحان شده با استفاده از طراحی ما:

			100		6 7 7 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1				
[2022-07-29	05:29:40 EDT]	iverilog	'-Wall'	'-g2012'	design.sv	testbench.sv	&& unbuffer	vvp a.out	
Matrix A:									
1	4								
2	4								
Matrix B:									
0	1								
1	1								
Mult Result:									
4	5								
4	6								



# نحوه برخورد با Floating point:

فایل های تحت عنوان جمع کننده و ضرب کننده خودکار اعداد اعشاری به ما داده شده بود که متاسفانه این فایل ها مشکلاتی داشتند. ابتدا از همه با توجه به ترکیبی بودن کل فرایند ضرب ماتریس در برنامه ما کلاک داشتن و ترتیبی بودن آن ها کار را خراب میکرد دوما به هیچ عنوان قابل سنتز نبودند و سوم ایرادات ریز سینتکسی داشتند که قادر به حل آن مشکلات در زمان کوتاه باقی مانده نبودیم فلذا مجبور شدیم خودمان یک جمع کننده و ضرب کننده اعداد اعشاری بالا بیاوریم ک آن ها را در دو ماژول ترکیبی جدا پیاده سازی کرده و با استفاده از وربلاتور و تست های CPP درستی آن ها را مشخص کردیم.

متاسفانه با توجه به زمان اندک باقی مانده تا ارائه نتوانتسیم تست خودکار برنامه را راه بیندازیم ولی با توجه به پاسخگویی های دستی برنامه و همچنین پاسخ های درست آن در استیت های قبلی و همچنین درستی تست ها میتوان از پاسخ ارسالی برنامه اطمینان حاصل کرد.