



**دانشگاه صنعتی امیرکبیر**  
**( پلی تکنیک تهران )**  
دانشکده مهندسی کامپیوتر

درس فناوری های حافظه  
دکتر حامد فربه

رضا آدینه پور  
۴۰۲۱۳۱۰۵۵

## تمرین شبیه سازی چهارم

تدریسار:

مرتضی عادلخانی ([madelkhani@aut.ac.ir](mailto:madelkhani@aut.ac.ir))  
سارا زمانی ([sara.zamani73@aut.ac.ir](mailto:sara.zamani73@aut.ac.ir))

## سوالات تئوری

۱. به سوالات زیر پاسخ دهید:

۱. PUM چیست و کدام نوع حافظه ها برای آن بیشتر استفاده می شوند؟ توضیح دهید چرا هر نوع حافظه استفاده می شود.

**پاسخ:** پردازش در حافظه (PUM) یک کانسپت محاسباتی است که در آن برخی از محاسبات ساده مانند جمع و ضرب به جای انتقال داده ها بین CPU و حافظه، مستقیماً در حافظه انجام می شوند. معمولاً از DRAM، SRAM و NVM ها در PUM استفاده می شود. که در ادامه به بررسی مزایا و معایب استفاده از هر کدام می پردازیم.

DRAM ها به دلیل اینکه رایج ترین نوع حافظه فرار با تراکم بالا و هزینه کم به ازای هر بیت هستند، به طور گسترده استفاده می شود. ویژگی های خازنی سلول های DRAM امکان انجام تکنیک های محاسباتی درون حافظه مانند عملیات منطقی و حسابی را فراهم می کند. مزایا: تراکم بالا، ارزان است.

معایب: فرار، نیاز به تازه سازی دوره ای و معمولاً تأخیر بیشتر نسبت به SRAM. اما در مقابل SRAM زمان های تأخیر کمتری و زمان دسترسی سریعتری نسبت به DRAM دارد و در مواردی که سرعت برای ما بسیار مهم است، (مانند Cache)، استفاده می شود. توانایی حفظ حالت بدون نیاز به تازه سازی، آن را برای عملیات های PUM مناسب می سازد.

مزایا: زمان های دسترسی سریع نسبت به DRAM، عدم نیاز به تازه سازی. معایب: تراکم کمتر و هزینه بیشتر به ازای هر بیت نسبت به DRAM.

در مقابل حافظه های فرار، انواع حافظه های غیر فرار مانند Flash، PCM و ReRAM به دلیل نگه داشتن داده بدون برق، برای ذخیره سازی پایدار و محاسبات مناسب هستند. این حافظه ها می توانند برخی عملیات منطقی را درون سلول های حافظه انجام دهند. مزایا: غیر فرار بودن.

معایب: عموماً سرعت نوشتن کندتر و دوام کمتر نسبت به DRAM و SRAM.

۲. نقاط ضعف UPMEM چیست؟

**پاسخ:** از نقاط ضعف UPMEM ها می توان به موارد زیر اشاره کرد:

(آ) انعطاف پذیری و قابلیت برنامه ریزی محدود:

معماری PIM UPMEM برای انواع خاصی از عملیات (مانند وظایف داده محور مانند جست و جو در پایگاه داده و تحلیل) است. ممکن است به اندازه CPU یا های GPU سستی عمومی و همه منظوره نباشد، که کاربرد آن را به بارهای کاری خاص محدود می کند.

(ب) یکپارچه سازی و سازگاری:

یکپارچه سازی ماژول های PIM UPMEM با سیستم های موجود می تواند چالش برانگیز باشد. ممکن است مشکلات سازگاری با معماری های حافظه و پردازنده فعلی به وجود بیاید که نیاز به اصلاحات در کانفیک نرم افزار و سخت افزار دارد.

(ج) مسائل مربوط به کارایی انرژی:

در حالی که PIM هدفش کاهش مصرف انرژی با حداقل کردن حرکت داده ها بین حافظه و CPU است، صرفه جویی واقعی در انرژی می تواند وابسته به بار کاری باشد. برخی عملیات ممکن است همچنان مصرف انرژی قابل توجهی داشته باشند، به خصوص اگر منطق PIM به طور کامل برای آن کاربرد به خصوص بهینه سازی نشده باشد.

(د) توسعه و اشکال زدایی:

همانطور که در کلاس هم بررسی شد، توسعه برنامه ها برای PIM نیاز به مدل های برنامه نویسی و

ابزارهای جدید دارد. دیباگ و پروفایل کردن برنامه های PIM می تواند به دلیل طبیعت توزیع شده و درون حافظه ای محاسبات سخت تر از برنامه نویسی CPU/GPU سنتی باشد.

۳. ساختار Ambit را معرفی کرده و مزایا و معایب آن را توضیح دهید.

**پاسخ:** Ambit یک معماری PIM است که از بستر DRAM موجود برای انجام عملیات بیتی (مانند NOT، OR، AND) به طور مستقیم درون حافظه استفاده می کند. این معماری از ویژگی های آنالوگ سلول های DRAM و Sense Amplifier برای اجرای این عملیات استفاده می کند. از مزایای آن می توان به موارد زیر اشاره کرد:

(آ) کاهش حرکت داده ها:

با انجام محاسبات مستقیماً درون DRAM به طور قابل توجهی نیاز به حرکت داده بین CPU و حافظه را کاهش می دهد، که منجر به کاهش تاخیر و مصرف انرژی می شود.

(ب) توان محاسباتی بالا:

Ambit می تواند عملیات بیتی را بر روی حجم زیادی از داده ها به طور همزمان انجام دهد، که توان محاسباتی بالایی برای وظایف داده محور مانند جستجوهای پایگاه داده، رمزنگاری و شبکه های عصبی فراهم می کند.

(ج) تغییرات سخت افزاری حداقلی:

Ambit از زیرساخت DRAM موجود با تغییرات حداقلی استفاده می کند، که ادغام آن را با سیستم های فعلی نسبت به معماری های PIM آسان تر می کند.

همچنین از معایب Ambit می توان به موارد زیر اشاره کرد:

(آ) محدودیت در انواع عملیات:

Ambit به طور عمده برای عملیات بیتی طراحی شده است. نمی تواند به طور کارآمد محاسبات پیچیده تر ریاضی یا اعشاری را انجام دهد، که کاربرد آن را به انواع خاصی از عملیات ها محدود می کند.

(ب) پیچیدگی در برنامه نویسی:

برنامه نویسی برای Ambit نیاز به درک مدل عملیاتی خاص و محدودیت های آن دارد. برنامه نویس ها باید الگوریتم های خود را برای استفاده موثر از عملیات بیتی تطبیق دهند، که می تواند پیچیدگی نرم افزاری را افزایش دهد.

(ج) چالش های مقیاس پذیری:

در حالی که Ambit توان محاسباتی بالایی برای عملیات بیتی ارائه می دهد، مقیاس پذیری آن به سیستم های حافظه بزرگ تر یا ادغام آن با واحدهای پردازشی دیگر ممکن است چالش هایی از نظر هماهنگی و مدیریت ایجاد کند.

## سوالات شبیه سازی

۲. در این بخش از تکلیف خود، شما با شبیه ساز MNSIM 2.0 برای پیاده سازی یک شتاب دهنده شبکه عصبی سر و کار خواهید داشت. بنابراین، ابتدا باید این شبیه ساز را دانلود کرده و نتایج را طبق درخواست ارائه دهید.

### پیکربندی پایه:

۱. پارامترهای شبکه عصبی VGG8 که بر روی مجموعه داده CIFAR-10 آموزش دیده است را دانلود کنید، همانطور که در راهنمای MNSIM 2.0 توضیح داده شده است.

۲. برای هر اجرا، باید VGG8 را به عنوان شبکه عصبی مورد نظر خود انتخاب کنید.

پاسخ: وزن ها را از اینجا دانلود می کنیم.

### سوال ۱:

همانطور که در کلاس یاد گرفتید، ساختار PIM شامل تعدادی کاشی است و هر کاشی شامل تعدادی PE است. در هر PE، ما مدارهای ضروری و ساختار ضربدری سلول های حافظه را داریم. اگر اندازه ضربدری را کاهش دهیم چه اتفاقی می افتد؟ به عنوان مثال، آیا باید انتظار داشته باشیم که توان، تأخیر و دقت کاهش، افزایش یا تغییر نکند؟ پاسخ خود را به تفصیل توضیح دهید.

### سوال ۲:

اگر بخواهیم PUM را به این شبیه ساز اضافه کنیم، کدام قسمت باید تغییر کند؟

### سوال ۳:

در اولین پیاده سازی، ابعاد ضربدری (Xbar) را به  $256 \times 256$  تنظیم کنید. در پیاده سازی دوم، ابعاد ضربدری را به  $128 \times 128$  تغییر دهید. مجموع تأخیر، توان و انرژی را گزارش دهید و جدول را پر کنید. چه اتفاقی افتاد؟ چرا؟ (برای هر پارامتر به تفصیل توضیح دهید.)

وضعیت	$256 \times 256$	$128 \times 128$	
تأخیر			کاهش / افزایش
توان			کاهش / افزایش
انرژی			کاهش / افزایش

جدول ۱: جدول I