بسمه‌تعالی

درخواست طرح پیشنهادی (RFP)

پژوهش، طراحی، نمونه‌سازی، تست و تدوین دانش فنی تراشه شتاب‌دهنده CNN مبتنی بر MAC Array

با فناوری ۴۰ نانو متر Mixed Signal

تیر ماه ۱۴۰۵

۱- مقدمه

در سال‌های اخیر با رشد بی‌سابقه نیاز به پردازش هوش مصنوعی، پردازش ابری، شبکه‌های پرسرعت و سامانه‌های ذخیره سازی داده، طراحی شتاب‌دهنده سخت‌افزاری به یکی از ارکان اصلی معماری پردازش‌های پیشرفته تبدیل شده است. در حالی که مدل‌های زبانی بزرگ (LLMها) بیشتر به معماری transformer متکی هستند، شبکه‌های CNN (Convolutional Neural Network) ستون اصلی پردازش تصویر، بینایی ماشین و تشخیص ویدیو (مانند الگوریتم‌های YOLO یا ResNet) به شمار می‌روند. شتاب‌دهنده‌هایی که برای اجرای الگوریتم‌های CNN طراحی می شود، به دلیل ساختار تکراری کانولوشن‌ها و نیاز به بازاستفاده مداوم از داده‌ها، از جریان‌های داده اختصاصی (مانند Weight-Stationary یا Row-Stationary) برای حذف گلوگاه‌های حافظه استفاده می کنند. یک شتاب‌دهنده تخصصی CNN دارای بخش‌های زیر است:

- بافرهای خطی (Line Buffers): رجیسترهای سخت‌افزاری که نوارهای افقی کوچکی از تصویر را نگه می‌دارند تا هنگام حرکت پنجره کانولوشن روی تصویر، نیازی به بارگذاری مجدد کل فریم نباشد؛
- واحد های Pooling سخت‌افزاری: بلوک های منطقی که عملیات Max-Pooling یا Average-Pooling را فوراً و بدون هدر دادن چرخه‌های کلاک پردازنده اصلی انجام می‌دهند؛
- مکانیزم حذف صفرها (Sparsity Engines): شبکه‌های CNN پس از لایه ReLU حاوی صفرهای زیادی هستند. این کنترلرها با تشخیص صفر، عملیات ریاضی را کاملاً نادیده می گیرند که این کار سرعت را دو برابر کرده و مصرف توان را نصف می کند.

از جمله مهم‌ترین شتاب‌دهنده‌های طراحی‌شده برای CNNها می‌توان به موارد زیر اشاره کرد:

۱- میکروکنترلرهای کم‌مصرف جاسازی شده برای استنتاج لبه (Edge IoT): این تراشه‌ها برای دوربین‌های هوشمند "همیشه روشن" و تجهیزات پوشیدنی طراحی شده‌اند تا مدل‌های بینایی ماشین را با یک باتری کوچک اجرا کنند:

- تراشه‌های MAX۷۸۰۰۲ / MAX۷۸۰۰۰ (ساخت Analog Devices): این میکروکنترلرها دارای یک شتاب‌دهنده سخت‌افزاری اختصاصی CNN با ۶۴ پردازنده موازی هستند. این لایه سخت‌افزاری، حلقه‌های درهم‌تنیده پردازشی (Nested Loops) را به صورت موازی اجرا کرده و با نگه‌داشتن تمام وزن‌ها روی خود تراشه، مصرف توان را به چند میلی‌وات کاهش می‌دهد.

۲- پردازنده‌های جاسازی شده بینایی در لبه (VPUs / NPUs): این کمک‌پردازنده‌ها (Co-processors) به کامپیوترهای لبه یا سیستم‌های صنعتی متصل می‌شوند تا فید چند دوربین را به صورت هم‌زمان و بی‌درنگ پردازش کنند:

- Hailo-۸ / Hailo-۱۰: این شتاب‌دهنده ساختاری منحصربه‌فرد دارد که در آن حافظه و عناصر مسیریابی مستقیماً در کنار عناصر پردازشی توزیع شده‌اند. این چیدمان، ساختار لایه‌به‌لایه‌ی شبکه‌های CNN را شبیه‌سازی کرده و نرخ فریم بسیار بالایی را با کمترین توان مصرفی ارائه می‌دهد؛
- Intel Movidius / Keem Bay: یک واحد پردازش بینایی (VPU) اختصاصی است که به خطوط استریم سخت‌افزاری مجهز شده تا عملیات کانولوشن دوبعدی را با سرعت بالا روی تصاویر اعمال کند.

۳- مقیاس دیتاسنتر و ابری:

- Google TPU (نسخه‌های اولیه): اگرچه نسخه‌های مدرن TPU گوگل اکنون برای مدل‌های زبانی (LLM) بزرگ شده‌اند، اما معماری نسخه‌های ۱ تا ۳ آن‌ها کاملاً بر پایه آرایه‌های سیستولیک (Systolic Array) برای شتاب‌دهی به جبر خطی و کانولوشن‌های CNN طراحی شده بود. در این ساختار، داده‌ها در یک شبکه دوبعدی از بلوک‌های ضرب-جمع (MAC) بدون نیاز به نوشتن و خواندن مداوم در حافظه خارجی به حرکت درمی‌آیند.

۴- بلوک‌های سخت‌افزاری قابل برنامه‌نویسی برای FPGA:

- AMD Xilinx DPU (Deep Learning Processor Unit): یک هسته IP قابل پیکربندی است که به طور ویژه برای CNN بهینه‌سازی شده و عملیات زمان‌بندی، ضرب ماتریسی موازی و لایه‌های Pooling را مستقیماً در منطق برنامه‌نویسی FPGA انجام می‌دهد؛
- پروژه Eyeriss (دانشگاه MIT): یک معماری پایه‌ای برای تراشه‌های مدرن CNN است که جریان داده Row-Stationary را معرفی کرد. این روش بازاستفاده از پیکسل ها و وزن ها را درون رجیسترهای محلی به حداکثر می‌رساند. کدهای متن‌باز VHDL/Verilog این پروژه برای کارهای تحقیقاتی در دسترس است.

۲- هدف فراخوان

هدف از این فراخوان، شناسایی و انتخاب تیم مناسب برای پژوهش، طراحی، نمونه‌سازی، تست و تدوین دانش فنی یک شتاب‌دهنده هوش مصنوعی سبک (Lightweight AI Accelerator) برای inference با تمرکز بر اجرای کارآمد شبکه‌های عصبی کانولوشنی (CNN) برای کاربردهای: بینایی ماشین (Object Detection, Classification)، رباتیک (tracking, navigation)، سیستم‌های نظارتی (surveillance AI)، IoT هوشمند، پردازش تصویر صنعتی و edge AI camera modules می‌باشد. الزامات عملکردی برای مدل‌های هدف که این پردازنده باید پشتیبانی کند شامل موارد زیر می‌باشد:

- MobileNetV۱ INT۸: حداقل ۳۰ FPS برای ورودی ۲۲۴×۲۲۴؛
- Tiny-YOLO: حداقل ۱۰ FPS برای ورودی ۴۱۶×۴۱۶ ؛
- ResNet-۱۸ (quantized): latency کمتر از ۵۰ ms؛
- MAC array utilization ≥ ۶۰٪ under defined benchmark workloads (ResNet-۱۸) measured over full inference execution including memory stalls.

شرایط آزمون:

Batch size = ۱
Input data from off-chip DRAM

تمامی اندازه‌گیری‌ها باید شامل data movement overhead باشند.

برای توسعه‌های آتی و نسخه‌های بعدی این پردازنده، لازم است که از معماری برپایهٔ ریزتراشه (chiplet) استفاده شود تا اولاً بتوان از یک بلوک / ریزتراشه در چندین تراشه استفاده کرد و بدین ترتیب تیراژ ساخت را افزایش داد تا هزینهٔ ساخت قابل توجیه‌تر باشد. همچنین امکان به روز رسانی یا توسعه بخشی از تراشه بدون طراحی و ساخت مجدد کل آن میسر باشد. علاوه بر این کارایی تراشه‌های مبتنی بر chiplet معمولاً بالاتر و بازده تولید (yield) آنها نیز بیشتر است و بازاریابی و فروش آنها راحت‌تر می‌باشد. برای نیل به این هدف در نسخه‌های آتی، ضروری است که این تراشه در این نسخه، هر چند بصورت chiplet، طراحی و ساخته نمی‌شود، لیکن بصورت chiplet ready طراحی شود تا تبدیل آن به chiplet در نسخه‌های آتی سهل‌تر باشد. الزامات معماری chiplet ready ذیلاً در فصول آتی به استحضار می‌رسد. طراحی هرگونه لایه فیزیکی خارجی (PHY)، از جمله DDR PHY یا PCIe PHY یا ... در دامنه این طرح قرار نمی‌گیرند و تبادل داده با حافظه خارجی برعهده CPU/SoC میزبان خواهد بود.

۳- مدل اجرای طرح

طرح باید در فرآیند mixed signal نود فناوری ۴۰ نانومتر اجرا شود. در همین راستا، فایل‌ها و اسناد لازم شامل PDK فناوری مذکور، توسط مرکز ملی پیشران طراحی تراشه در اختیار مجری قرار خواهد گرفت و پشتیبانی‌های فنی و نرم‌افزاری از تیم مجری به عمل خواهد آمد. پس از اتمام طراحی، فایل‌های GDSII مربوطه توسط مرکز ملی پیشران طراحی تراشه و با حمایت برنامه ملی میکروالکترونیک، برای ساخت (Tape Out) به کارخانهٔ تولید کنندهٔ تراشه (Foundry) ارسال خواهد شد و مجری نیازی به تعامل شخصی با Foundry نخواهد داشت. پس از ساخت تراشه، عملیات باندینگ توسط مرکز ملی پیشران طراحی تراشه و در نهایت تست و ارزیابی آن مشترکاً توسط مجری و مرکز ملی پیشران طراحی تراشه انجام خواهد گردید.

۴- محدوده و دامنه طرح‌های پیشنهادی

معماری سیستم علاوه بر این تراشه شامل یک host SoC / CPU مستقل است که وظیفه مدیریت حافظه خارجی، بارگذاری مدل‌ها، زمانبندی تسکها و ارتباط کلی با Accelerator را بر عهده دارد. Accelerator بصورت یک واحد محاسبات تخصصی عمل کرده و از طریق یک رابط استاندارد با میزبان ارتباط برقرار خواهد کرد.

مساحت مورد نیاز برای طراحی ترجیحاً باید کمتر از ۴ میلیمتر مربع و الزاماً کمتر از ۹ میلیمتر مربع باشد و طرح‌های پیشنهادی باید موارد زیر را دربر گیرند:

- **طراحی RTL کامل**
- **توسعه نرم‌افزار پایه**
- **مشخصات کلی تراشه**

- Technology Node: ٤٠nm
- Frequency Target: ٢٠٠–٤٠٠ MHz
- Compute Precision: INT٨ (primary)
- Accumulator: INT٣٢
- Peak Throughput: ~٠٫٥–١ TOPS
- On-chip SRAM: ١٢٨–٢٥٦ KB
- Power Budget: <١W
- Die Size Target: ٢–٥ mm²

معماری داده (Dataflow)

- Row-Stationary (ترجیحی)؛
- Weight-Stationary (در صورت توجیه PPA).

ارائه تحلیل مقایسه‌ای (energy vs reuse)

مشخصات بلوک‌های سخت‌افزاری تراشه

- MAC Array (Core Block)
  Type: Systolic Array
  Size: ۳۲×۳۲
  Precision: INT۸×INT۸ to INT۳۲
  Features:
    - pipeline
    - clock gating
    - configurable tiling
    - local reuse support

  Additional Requirements:
    - systolic dataflow must be explicitly defined (input/weight/output flow direction)
    - pipeline depth per MAC stage must be determined
    - support for partial sum accumulation across tiles
    - possibility of stall-free streaming if data is provided
    - inter-PE communication latency ≤ ۱ cycle
    - The memory subsystem must support sufficient bandwidth to avoid MAC starvation.
    - Designers must provide bandwidth analysis demonstrating no underutilization due to memory bottlenecks.
- Buffer Subsystem
  Buffers:
    - Input Buffer
    - Weight Buffer
    - Output Buffer

  Features:
    - double buffering (ping-pong)
    - banked SRAM
    - parallel access
- SRAM
  SRAM should be implemented as multi-bank single-port or dual-port memory.
  logical multi-port access must be achieved via banking and arbitration.
  bank conflict handling must be determined.

Total: ۱۲۸–۲۵۶ KB

Type: multi-bank SRAM

SRAM latency target: ≤ ٤ cycles, design must specify read/write latency and banking strategy explicitly.

Support:

- simultaneous read/write
- low latency

- DMA Engine

  Channels: ≥۲

  Features:

  - burst transfer (AXI٤)
  - address generation
  - configurable stride
  - interrupt (optional)
  - support for memory-mapped and streaming modes
  - descriptor-based operation (optional but preferred)
  - latency hiding via prefetch
  - back-pressure handling from compute unit

- AXI Interface

  AXI٤ (data path)

  AXI-Lite (control)

  Features:

  - burst support
  - high throughput
  - AXI Interface Requirements:
  - AXI٤ data width: حداقل ۶۴ یا ۱۲۸ بیت
  - burst length: حداقل ۱۶
  - target bandwidth: ۴–۲ حداقل GB/s
  - QoS considerations to prevent starvation

- Control Unit

  Type: FSM + Loop Engine

  A clear programming model must be defined:

  - Instruction-based
  - Register-based
  - Microcode-based

  The choice must be justified in terms of flexibility vs hardware overhead.

  Capabilities:

  - nested loops (٦D):
    - ✓ channel in/out
    - ✓ spatial
    - ✓ kernel
  - programmable via registers
  - Control Unit must support:

- ✓ programmable instruction sequence یا microcode
- ✓ loop nesting تا حداقل ۶ سطح
- ✓ runtime reconfiguration without reset

- Activation Unit
  - ReLU (mandatory)
  - optional: clamp / leaky ReLU
- Quantization Unit
  - INT۳۲ → INT۸
  - scaling + shift
  - support for per-channel quantization (ترجیحی)
  - rounding mode must be determined (nearest / stochastic)
  - overflow handling: saturation (mandatory)
- PMU (Performance Monitoring Unit)
  - cycle counter
  - MAC utilization
  - stall tracking
  - bandwidth usage
- DFT
  - scan chain
  - MBIST for SRAM
  - test mode control

هر یک از بخش‌های فوق باید با ارائه deliverableهای مشـــخص (مطابق بخش ۷) و گزارش‌های قابل ارزیابی تکمیل شوند. صرف انجام فعالیت بدون ارائه مستندات و نتایج قابل اندازه‌گیری قابل قبول نیست.

۵- قابلیت چیپلتی داشتن (Chiplet-Ready Requirements)

به منظور قابل تقسیم بودن تراشه طراحی‌شده به چند چیپلت، در طرح‌های پیشنهادی رعایت موارد زیر ضروری است:

[redacted]شرایط طراحی

- partition-based design
- clean interface boundary
- decoupled subsystems

[redacted]راهبرد ارتباطی

Internal:

- AXI٤
- AXI-Stream (optional)

Future-ready:

- Transport Layer (AXI to packet abstraction)
- Adapter placeholder for UCIe PHY
- latency budget for internal interface should be determined
- packetization boundary definition (transaction layer abstraction)
- clock domain crossing strategy must be defined
- die-to-die can be written to UCIe protocol stack (PHY-independent design)
- interface should be designed in a manner that without changing the compute logic in the next versions we can add:
    - CHI adapter;
    - UCIe PHY.

[redacted]تقسیم‌بندی پیشنهادی

Chiplet A: Compute (MAC + SRAM);
Chiplet B: IO + DMA + Host Interface.

۶- نرم‌افزارهای نیازمند توسعه

[redacted]راه‌انداز (driver)

Language: C
Features:

- register configuration;
- DMA control;
- execution control.

[redacted]کتابخانه زمان اجرا (runtime library)

APIs:

- conv۲d();
- load_weights();
- run_layer().

[redacted]مبدل مدل (model convertor)

Language: Python
Quantization accuracy loss must be reported: Top-۱ accuracy degradation ≤ ۳٪ نسبت به مدل FP۳۲
Functions:

- PyTorch → INT۸
- weight extraction
- config generation

Model Converter must support:

- ONNX input (mandatory)
- PyTorch export
- graph optimization (layer fusion)
- memory scheduling (tiling generation)

نرم‌افزار آزمون و ارزیابی (test and evaluation software)

- golden model (NumPy);
- bit-accurate comparison;
- validation scripts.

۷- شرایط و الزامات تست، ارزیابی و تأیید

تأیید RTL (RTL Verification)

- unit-level test bench;
- system-level simulation;
- random + directed tests.

تأیید عملکردی (Functional Validation)

- CNN inference correctness;
- comparison with PyTorch.

تأیید کارآیی (performance validation)

- throughput measurement;
- utilization analysis.

تأیید بعد از ساخت (post silicon validation)

- bring-up:
  - JTAG debug interface (mandatory);
  - Possibility of complete register read back;
  - internal signal observability (via scan یا debug mux).
- hardware testing;
- real model execution.

معیارهای موفقیت این طرح عبارت خواهند بود از:

- اجرای صحیح CNN ؛
- دستیابی به throughput هدف؛
- tape-out موفق؛

- silicon functional:
  - Functional Coverage ≥ ۸۰٪
  - Code Coverage ≥ ۸۰٪
  - assertion-based verification
  - formal verification for vital blocks (optional but preferred)

۸- خروجی‌ها و موارد تحویلی مورد انتظار

[redacted]خروجی‌های شبیه‌سازی

الف) RTL کامل و تست‌شده:

- کد RTL به‌صورت Synthesizable Verilog/System Verilog؛
- رعایت Coding Style استاندارد؛
- پوشش Functional Coverage حداقل ۸۰٪.

ب) محیط کامل شبیه‌سازی:

- تست‌بنچ‌ها (UVM یا معادل آن)؛
- تست‌های Functional ، Corner Case، Throughput، Latency؛
- مدل‌های رفتار(Behavioral Models) ؛
- اسکریپت‌های شبیه‌سازی (Questa/VCS/Xcelium) .

ج) گزارش PPA در فناوری ۴۰ نانومتر:

- گزارش فرکانس، توان، سطح اشغال؛
- نتیجه سنتز از Synopsys/Cadence ؛
- تحلیل Timing (STA) کامل.

[redacted]نرم‌افزارها

- Driver;
- Runtime library;
- Model converter.

[redacted]ملزومات نمونه‌سازی

الف) GDSII نهایی:

- فایل GDSII به همراه DRC clean و LVS clean؛

- DRC/LVS Report کامل.

ب) فایل‌های ماسک و Sign-off :

- LEF/DEF;
- Layer map;
- Extraction (PEX);
- Timing sign-off پس از PEX .

ج) مستندات لازم برای نمونه‌سازی (MPW) :

- فرم‌های PDK-specific ؛
- پکیج تست (Test Chip Package Plan) ؛
- لیست ریسک‌ها و نکات ساخت؛
- Liberty files (timing models);
- SDF back-annotated;
- power intent (UPF/CPF);
- IR drop analysis report;
- EM analysis report.

[redacted] ملزومات باندینگ

اندازه، موقعیت، گام و نقشه‌ی چیدمان تمام باندپدها، نقشه اتصالات (Bonding Diagram) و ... جز مواردی است که طراح باید پس از اتمام طراحی در اختیار مرکز ملی پیشران طراحی تراشه قرار دهد.

[redacted] ملزومات تست و ارزیابی

پس از تحویل دای، تیم باید موارد زیر را ارائه کند:

الف) برد تست (Test Board) :

- طراحی شماتیک؛
- فایل PCB؛
- نرمافزارهای کنترل تراشه و آزمون آن؛
- Interface مناسب.

ب) برنامه تست سیلیکون:

- پارامترهای Throughput، Latency، Power؛
- Functional test با بورد تست؛

- تست Corner (ولتاژ و دما) در حد امکانات آزمایشگاهی و بدون الزام به qualification صنعتی؛
- ثبت تمام باگ‌ها و نتایج تست.

ج) گزارش کامل Silicon Validation :

- مقایسه عملکرد واقعی با پیش‌بینی PPA ؛
- تحلیل اختلافات؛
- پیشنهاد برای Tape-out بعدی.

فایل‌های مستندسازی و تدوین دانش فنی

الف) مستند معماری:

- توصیف کامل AI accelerator ؛
- بلوک دیاگرام‌ها؛
- جریان داده و محاسبات؛
- Interface diagrams.

ب) دستورالعمل استفاده و کاربری:

- نحوه راه‌اندازی؛
- ریجستر مپ؛
- ساختار دستورات؛
- محدودیت‌ها.

پ) کد متن نرم‌افزار (Source Code).

۹- زمان‌بندی اجرای طرح

برای اجرای طرح ۱۸ ماه زمان پیش‌بینی شده است:

- مدت زمان لازم برای طراحی حداکثر ۱۰ ماه:
  - ✓ معماری (۱ ماه): شامل spec definition و architecture design؛
  - ✓ طراحی RTL و verification (test bench and integration) (۷ ماه): شامل MAC, DMA, AXI, control؛
  - ✓ طراحی فیزیکی (۲ماه): شامل synthesis و P&R و timing closure.
- نمونه‌سازی (Tape out): ۶ ماه؛
- باندینگ، تست و ارزیابی و گزارش‌نویسی: ۲ ماه.

۱۰- حمایت‌ها

کمک‌هزینه طراحی

کمک هزینه نیروی انسانی جهت پژوهش و طراحی تا سقف ۵۰ ملیارد ریال و طبق دستورالعمل بنیاد علم ایران در سال ۱۴۰۵ به شرح زیر خواهد بود:

- هیئت علمی: ۵۰ میلیون تومان در ماه (مجموع بودجه اعضای هیئت علمی در طرح نباید بیش از ۲۰ درصد بودجه طرح باشد)؛
- دانشجو پسا دکتری: ۳۰ میلیون تومان در ماه؛
- دانشجو دکتری: ۱۵ میلیون تومان در ماه؛
- دانشجو کارشناسی‌ارشد: ۸ میلیون تومان در ماه.

خدمات نمونه‌سازی (Tape out)

ارائه خدمات دسترسی به کارخانه ساخت تراشه (Foundry) و دریافت تکنولوژی‌فایل (PDK) جهت ایجاد فایل GDS II و نمونه‌سازی تراشه‌ها از طریق مرکز ملی پیشران طراحی تراشه و با حمایت برنامهٔ ملی میکروالکترونیک معاونت علمی، فناوری و اقتصاد دانش‌بنیان خواهد بود.

خدمات باندینگ و تست

این خدمات نیز با حمایت برنامهٔ ملی میکروالکترونیک معاونت علمی، فناوری و اقتصاد دانش‌بنیان و از طریق مرکز ملی پیشران طراحی تراشه ارائه خواهد شد.

۱۱- مالکیت معنوی

مالکیت کلیه دستاوردها و خروجی‌های فنی و علمی پروژه (شامل کدهای HDL، بلوک‌های IP، طرح فیزیکی GDSII، اسرار فنی، مستندات و ...) متعلق به طراح خواهد بود، لکن مرکز ملی پیشران طراحی تراشه اجازه بهره‌برداری غیرانحصاری، دائمی، غیرقابل‌فسخ و رایگان از تمامی این دستاوردها و خروجی‌ها را خواهد داشت و می‌تواند از آن‌ها در پروژه‌های دیگر استفاده نماید.