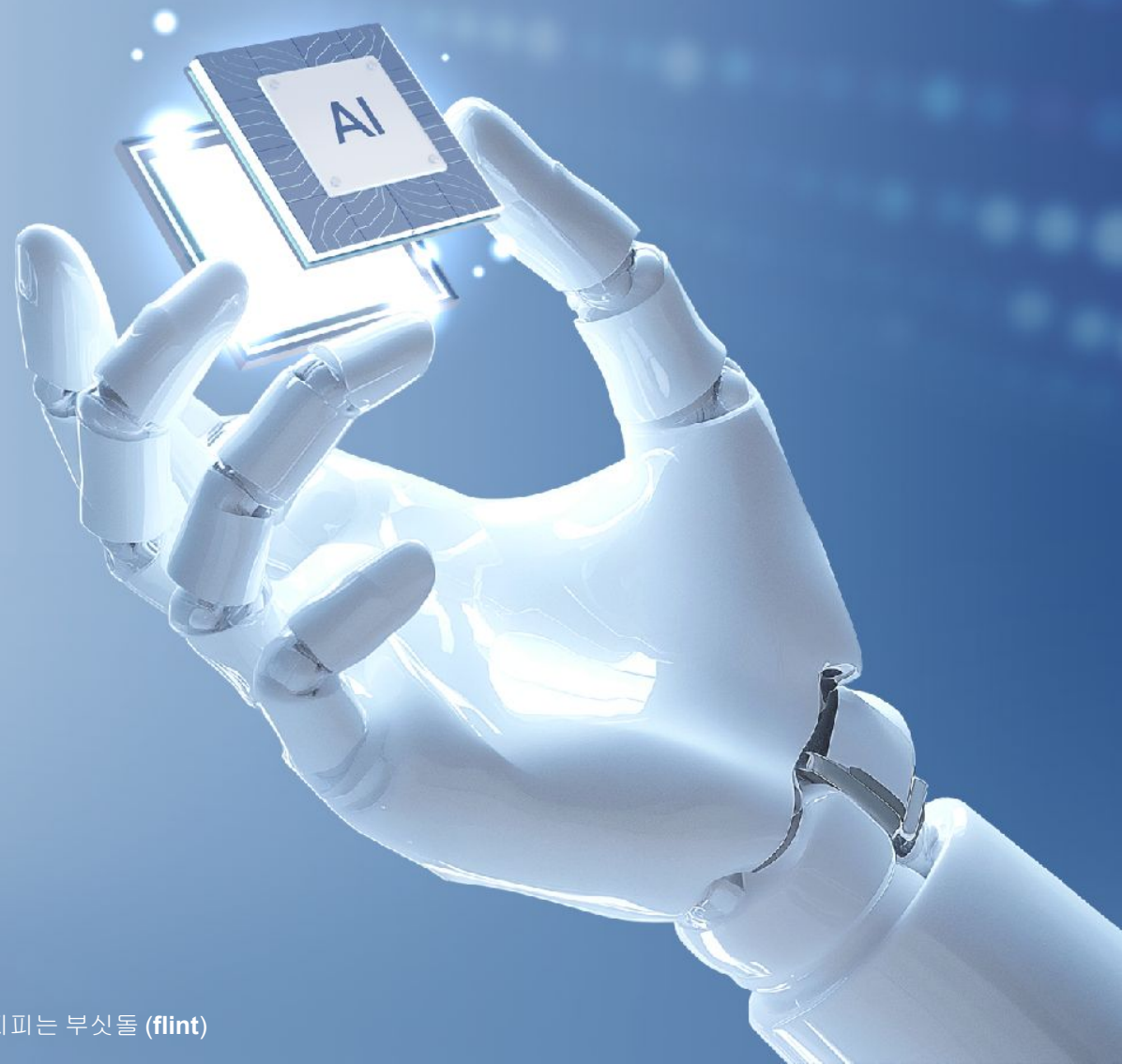
Confidential

hypervisual ai



The FLINT ® GPNPU Processor

CUDA와 동일한 개발환경을 지원하는 유연하고 효율적인 엣지 AI 프로세서

고성능 · 저전력 GPNPU로 온디바이스 AI 시장을 선도합니다.

정삼윤 · CEO & FOUNDER

2026

FLoating point 와 **INT**eger 연산이 가능한 Edge AI 프로세서 / **FL**exible & **INT**elligent / Edge AI 혁신의 불을 지피는 부싯돌 (**flint**)

혁신적인 GPNPU 아키텍처로 피지컬 AI 시대를 선도합니다.

COMPANY FACTS

설립	2023년 9월
소재지	서울시 강남구
제품	FLINT GPNPU IP · SoC · Module
타겟 시장	Physical AI 로봇 · 드론 · 자율주행 · CCTV
임직원	10명 핵심인력 반도체 · AI 25년+ 경력

정부 R&D / 지원사업 선정

2025	초격차 스타트업 **1000+**
2025	중기부 **TIPS R&D** 과제
2026	민관공동기술개발 **TRL 점프업**
2026	온디바이스 AI 반도체 **해외실증 사업**

4개 정부 과제 연속 선정
기술성 · 사업성 인정

투자 이력

SEED · 2025.08

6억원

Post Value 55억원 ·
케이그라운드벤처스,
킹고스프링

전략적 투자 (SI) · 2026.08.26

1억원

Post Value 97억원 ·
더인벤션랩 (아브락사스)

12개월 만에 기업가치 1.76× 상승

시장의 니즈와 괴리 □ 개발자 외면 □ 결국 CUDA로 복귀

엣지 NPU가 외면받는 이유 PRODUCT–MARKET MISFIT

CUDA와 호환되지 않는 SDK	• **파편화된 SDK**와 빈약한 커뮤니티 • 모델 포팅, 디버깅 난이도 가중 • **개발 비용과 시간이 급증**
INT8 · 행렬곱 위주의 고정 구조 (H/W 아키텍처 한계)	• **최신 피지컬 AI 모델 대응 한계** • 모델 업그레이드 제약 = **제품 유효수명 단축** • 연산 Fallback, **DRAM병목** 데이터 이동비용 급증
모델 경량화 시 정확도 하락	• FP32 □ INT8로 양자화로 **정확도 하락** • 피지컬 AI에서는 1%의 정확도 하락도 치명적

NVIDIA Jetson이 장악한 이유 DE FACTO STANDARD

극도의 연산 유연성	• H/W : GPU + Tensor Core 구조 (정수 · 실수 연산) • S/W : 모든 연산 대응 가능한 프로그래밍 유연성
독보적 CUDA 생태계	• cuBLAS · cuDNN 등 업계 표준 라이브러리 • 풍부한 예제 · 커뮤니티 · 기술지원
검증된 솔루션	• 로봇 · 드론 · 자율주행 시장 장악, 조 단위 매출

BUT, JETSON의 구조적 제약 = 당사의 진입 지점

가격 · 전력소모 · 발열 · 폼팩터 · 공급망 종속 이슈 >>

Jetson으로 다 cover하지 못하는 시장의 보완재!

hvai

CUDA와 동일한 개발환경을 지원하는 유연하고 효율적인 차세대 엣지 AI 프로세서 개발!

A 연산 유연성 및 높은 정확도

- INT8 ~ FP32 지원
- 스칼라 · 벡터 · 매트릭스 · 비선형 연산 대응
- 확장 가능한 칩 구조

B 전력, 연산효율, 가격

- DRAM Access 최소화
- Processor Utilization 향상
- 면적당 성능 (PPA) 개선

C 개발자 편의 제공

- 자체생태계구축 X / 업계 표준 생태계 지원 전략
- C, C++ 프로그래밍 가능한 구조

GPU + NPU 하이브리드 구조

GENERAL PARALLEL
GPGPU

FIXED FUNCTION
NPU

⇅ Real-time Workload Balancer ⇅

Shared Cache · Same Die

Direct Fusion Cache™

GPGPU | NPU

Shared SRAM · Zero-Copy Cache

Global Cache / Registers · Weight Buffer

↔ DRAM 왕복 없이 캐시 레벨에서 데이터 공유

CUDA 동일 S/W 환경

User Code — C++ / Python

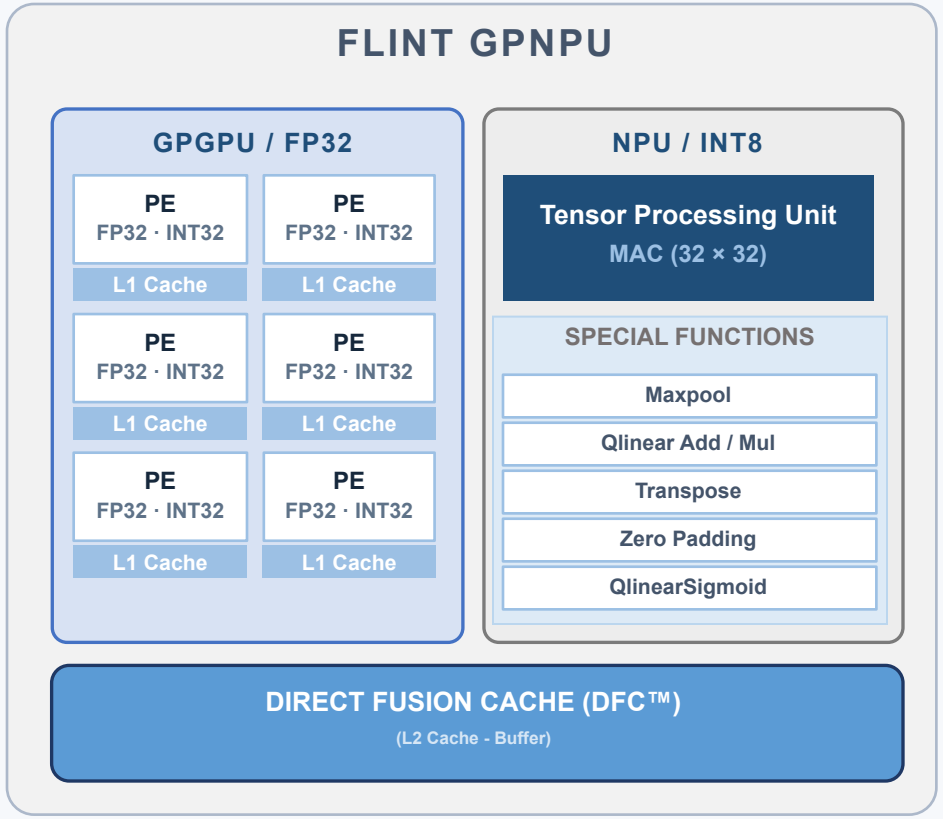
FUSE Compiler (Unified)

cuDNN | cuBLAS

FLINT GPNPU Core

피지컬 AI 최적의 하이브리드 아키텍처

GPGPU + NPU + Direct Fusion Cache (DFC™)



하이브리드 연산 처리

- GPGPU의 **유연성** 과 NPU의 **효율성** 을 통합.
- **INT8 ~ FP32, 스칼라 , 벡터 , 매트릭스 , 비선형 연산** 완벽 지원
- NPU가 고효율 INT8을, GPGPU가 고정밀 FP32(Softmax · 정규화)를 담당
- 모델 **경량화와 정확도** 동시에 최적화.

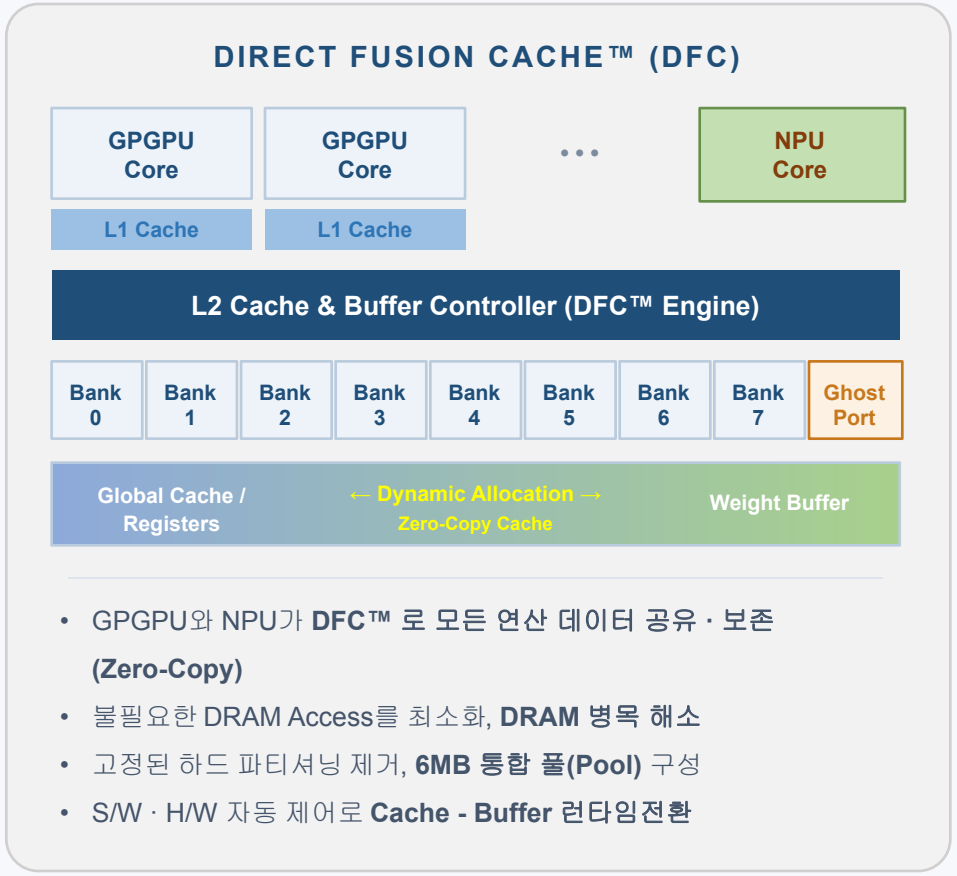
연산 할당 재구성 가능

- Workload Balancer : 실시간 **유닛별 부하를 조정** .
- GPGPU와 NPU 연산을 유동적으로 할당, **최적의 프로세싱 성능** 확보

확장 가능한 구조

- IoT (1 TOPS) ~ 엣지 서버 (40 TOPS) 까지 대응가능한 **확장형 구조** .

GPNPU 효율성의 핵심 (Direct Fusion Cache, DFC™)



- GPGPU와 NPU가 **DFC™ 로 모든 연산 데이터 공유 · 보존 (Zero-Copy)**
- 불필요한 DRAM Access를 최소화, **DRAM 병목 해소**
- 고정된 하드 파티셔닝 제거, **6MB 통합 풀(Pool)** 구성
- S/W · H/W 자동 제어로 **Cache - Buffer 런타임전환**

DATA 이동비용 최소화

DRAM-Bypass
- 시스템 대역폭 절약
- 전력 소모의 획기적 감소
- DRAM 병목에 따른 연산 지연 최소화

CHIP 면적대비 효율 향상

SRAM 파편화 제거
- GPU / NPU 메모리 **6MB 풀**로 통합, 동적 배분
- 유효 면적 대비 실효 캐시 극대화로 PPA 효율 향상

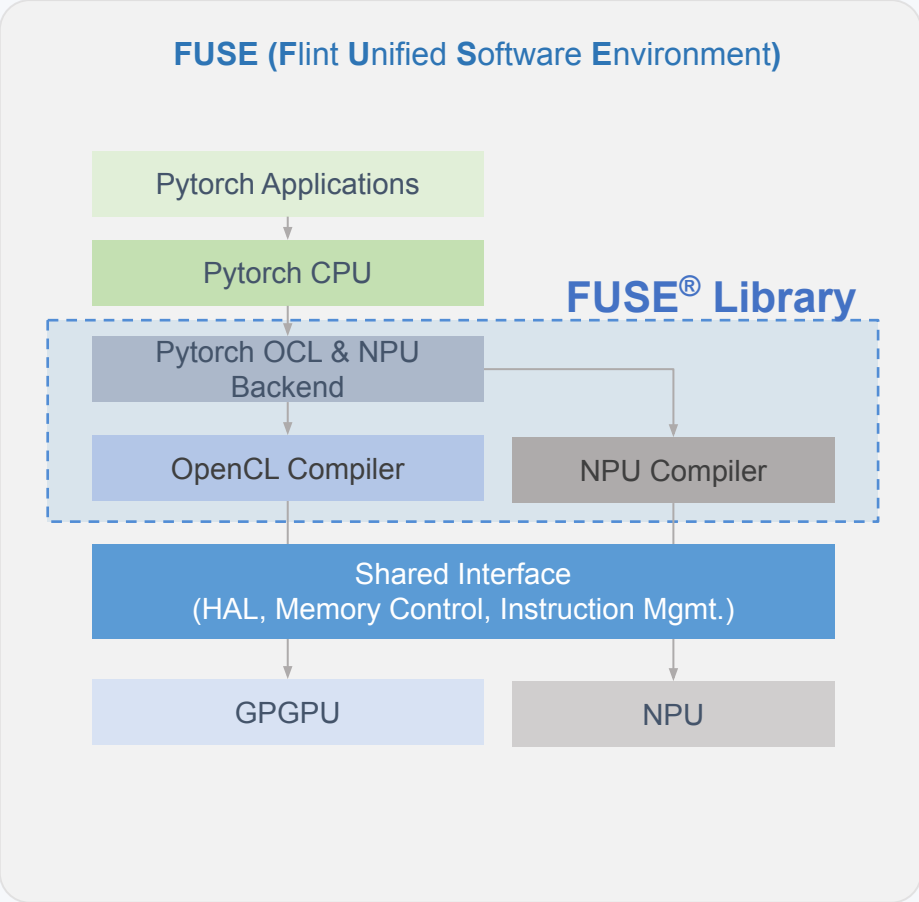
프로세서 Utilization 향상

Non-blocking & Prefetching
- NPU가 N번째 레이어 연산 중, 캐시 컨트롤러가 N+1번째 가중치를 미리 로딩(Prefetch)
- **유닛 대기시간(Stall / Idle) '0' 수렴**

특허 출원 3건 · 스케줄러 / 하이브리드 온칩 메모리 / 듀얼 버퍼링

CUDA Code 그대로 사용 - 학습곡선 “0”

자체 생태계를 새로 만들지 않고, 기존 CUDA 환경 지원 전략.



CUDA와 동일한 Software 개발환경

- **소스 코드 수정 없이 모델 포팅**
- NVIDIA 헤더 및 라이브러리를 GPNPU 환경으로 매핑
- **cuDNN · cuBLAS** 함수 라이브러리를 그대로 사용
- Jetson에서 동작하는 C / C++ · PyTorch 코드가 그대로 동작 - 즉각적 호환성
- 모델 포팅이 **Ctrl+C → Ctrl+V**로 단순화. 마이그레이션 비용 최소화 (Zero Refactoring)

VALUE 01

NVIDIA 종속성 탈피 (Lock-in 해제)

Jetson 높은 비용 · 소비전력 · 공급망 종속

□ **대체 기술 부재로 종속 심화**

소프트웨어 구조 변경 없이 1:1 하드웨어 교체가 가능

□ 기술 전환 리소스를 최소화

VALUE 02

제품 출시 기간 최적화

신규 NPU 도입 시 컴파일러 · 툴체인 **학습과 디버깅**에 수개월 규모의 개발 지연이 발생

업계 표준 cuDNN / cuBLAS 라이브러리를 지원

□ 학습곡선을 완화하고 TTM을 획기적으로 단축

hvai

GPU의 유연성과 NPU의 효율성을 동시에 확보한 유일한 아키텍처

비교 항목	**GPU** (NVIDIA Jetson)	**NPU** (Edge TPU / Hailo)	**GPNPU** (FLINT)
프로그래밍 유연성	✓ 자유로운 커스텀 커널	✗ 매트릭스 연산에 제한	✓ AI 특화 프로그래머블 C++ / Python 커스텀 지원
전력 · 발열 · 가격 · 사이즈	✗ 높은 전력 · 발열 · 사이즈 · 가격	✓ 최적화 모델에 효율적	✓ 면적당 성능 우위 Edge 폼팩터 대응
개발환경 (SW Stack)	✓ CUDA · cuDNN · cuBLAS	✗ 전용 SDK · OpenCL	✓ cuDNN / cuBLAS 호환 Zero Porting Cost
메모리 아키텍처	In-Core SIMT	Stand-alone Spatial	✓ Zero-Copy Heterogeneous Direct Fusion Cache™
Future-proof 대응	✓ 프로그래머블	✗ 새 연산자마다 재설계	✓ 하드웨어 교체 없이 대응 모델 진화에 유연

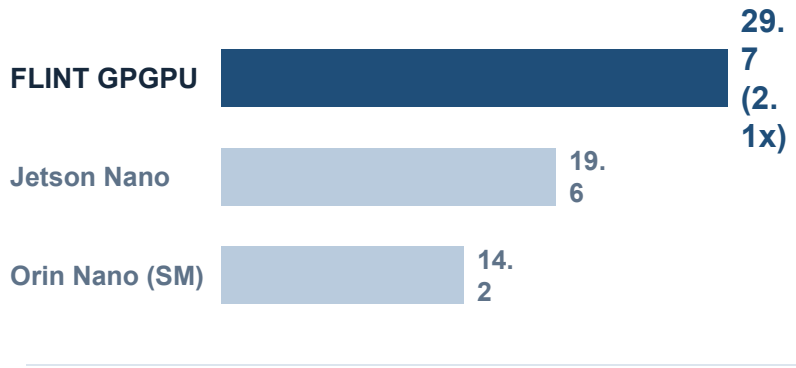
Flexible & Massive
성능, 유연성 우수
전력 · 가격 · 폼팩터에서 엣지 시장 적용 제약

Fixed & Efficient
최적화된 모델에서는 높은 효율
모델 변화 대응 · CUDA 생태계 단절로 개발자 이탈

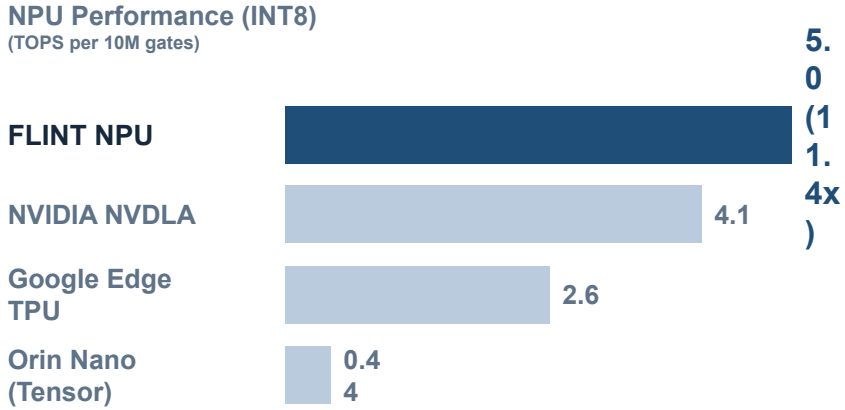
Flexible & Balanced ★
GPU 유연성 + NPU 효율성 + CUDA 호환
엣지 AI 최적 솔루션

면적당 성능 GPU 2.1× · NPU 11.4× vs Jetson Orin Nano

RTL Sim. 10M Gate Count당 연산 성능 비교우위



분류	아키텍처 / 공정	추정 면적	공식 성능	면적 효율
당사 제품	**FLINT GPGPU** 14nm / 1.0 GHz	약 10M	29.7 GFLOPS	🔥 **29.7 G**
Edge GPU	**Jetson Nano** 20nm / 921 MHz	약 120M	236 GFLOPS	약 19.6 GFLOPS
Edge GPU	**Orin Nano (SM)** 8nm / 625 MHz	약 900M	1,280 GFLOPS	약 14.2 GFLOPS



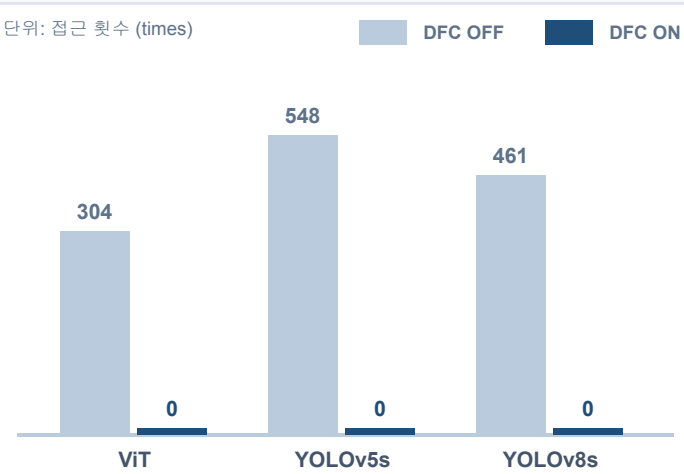
분류	아키텍처 / 공정	추정 면적	공식 성능	면적 효율
당사 제품	**FLINT NPU** 14/28nm (Dedicated)	약 8M	4.0 TOPS	🔥 **5.0 TOPS**
상용 NPU	**Google Edge TPU**	약 15M	4.0 TOPS	약 2.6 TOPS
상용 NPU	**NVIDIA NVDLA**	약 12M	5.0 TOPS	약 4.1 TOPS
통합 GPU	**Orin Nano (Tensor)**	약 900M	40.0 TOPS	약 0.44 TOPS

FLINT GPGPU 14nm 4-Core 1.0GHz, alupeak_absolute_fma 커널 RTL Sim 기준.
비교군: NVIDIA Jetson Orin Nano Series Data Sheet (1,280 GFLOPS FP32 / 40 TOPS INT8).

Direct Fusion Cache 효과

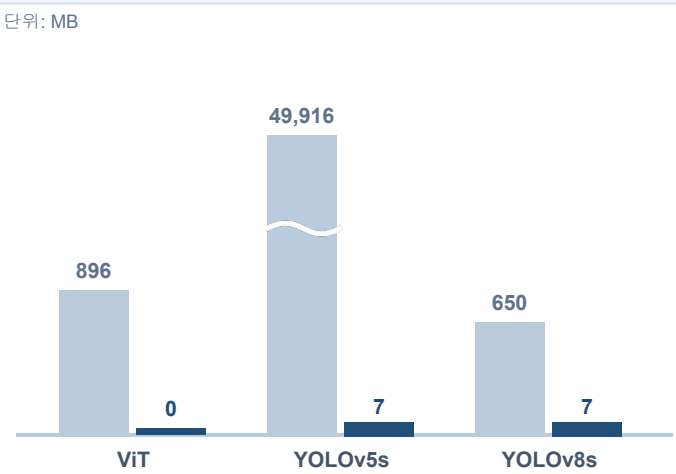
DRAM Access 100% 제거 · 대역폭 트래픽 99.6% 절감 · 추론 성능 평균 33배 향상.

DRAM ACCESS · 100% ↓



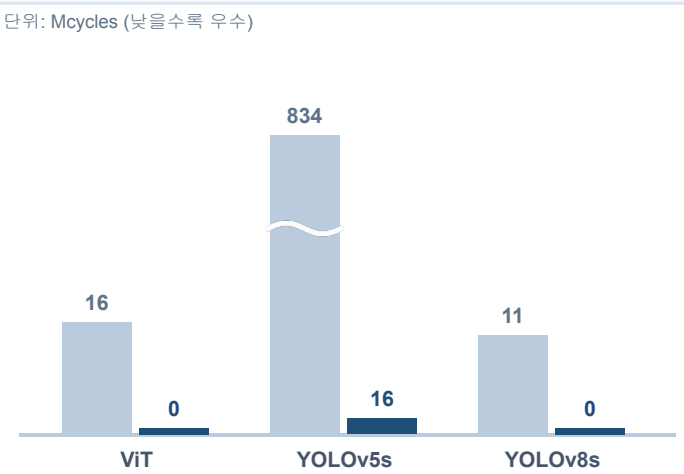
3개 모델 모두 외부 DRAM 접근 **0회** — 전량 On-Chip 처리.

BANDWIDTH TRAFFIC · 99.6% ↓



YOLOv5s 기준 **49,916 MB → 7 MB**, ViT는 0 MB.

INFERENCE CYCLE · 33× ↑



Memory Stall 제거로 유닛 대기시간(Idle) **0에 수렴**.

- DRAM 시스템 대역폭 클럭당 64 Bytes (200MHz 기준 12.8 GB/s)
- DRAM Read/Write Latency 100 Cycles + (Tensor Size / 64 Bytes/Cycle)
- Shared SRAM 클럭당 Direct Access (2 Cycles)

GPU 아키텍처 · AI 모델 · 사업화 전문가로 이루어진 팀.



CEO & FOUNDER

정삼윤

경영 총괄

- 한국시스템반도체협동조합 감사
- 전) 실리콘아츠 부사장
- SK · Kakao · Amazon · T-Mobile AI 디바이스 PM
- (주)하젠 창업 (등기이사)
- 삼성 해외영업 · 마케팅 · 사업기획
- 한국항공대학교 기계공학 학사



CTO & FOUNDER

윤형민 Ph.D.

인공지능 반도체 총괄

- 컴퓨터 아키텍처 분야 25년 경력
- 펩리스산업협회 감사
- 실리콘아츠 대표이사
- 삼성 종합기술원 프로젝트 리더
- 연세대학교 컴퓨터과학 박사



CAIO & FOUNDER

이종호 Ph.D.

인공지능 모델 총괄

- 서울대학교 빅데이터 혁신융합대학사업단 겸임교수
- 전) 가천대학교 반도체학부 AI 교수
- 롯데정보통신 인공지능 부문 상무
- 삼성 종합기술원 프로젝트 리더
- 서울대학교 컴퓨터공학 (AI) · 인지과학 박사

아브락사스 전략적 투자 참여

경영자문, 사업화,
투자유치 지원

전동수



한국공학한림원 정회원
前 삼성전자 반도체 메모리사업부장, 사장

학력
- 경북대학교 전자공학과 학사
- 경북대학교 전자공학과 석사

주요이력
- **前** 삼성전자 **CE**부문 의료기기사업부장 사장 (삼성메디슨 대표이사 겸임)
- **前** 삼성**SDS** 대표이사 사장
- **前** 삼성전자 **DS**사업총괄 메모리사업부장
- **前** 삼성전자 반도체사업부 메모리담당 사장
- **前** 한국반도체산업협회 제**9**대 회장

주요 외부 이력
- 한국공학한림원 정회원,
- 삼성그룹 기술대상 (2차례 수상),
- 제 28회 발명의 날 국무총리상 수상,
- 금탑산업훈장 (한국반도체산업협회)

조재문

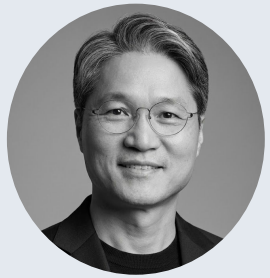


삼성 **Fellow**, 공학박사(Ph.D)
前 삼성전자 의료기기사업부 부사장

학력
- **KAIST** 전기및전자공학과 석사
- **KAIST** 전기및전자공학과 박사

주요이력
- **前** 삼성전자 의료기기사업부 부사장 (삼성메디슨 연구소장 겸직)
- **現 Nexium** 대표
- **現 First Step Labs** 대표
- 고려대기술지주회사 **(Mentor)**,
- **Techstars (Mentor)**,
- **BA Partners** (투자심의자문위원장),
- **First Step Labs** (엑셀러레이터 대표)

소병세



공학박사(Ph. D)
前 삼성전자 메모리사업부 부사장

학력
- 서울대학교 전자공학과 학사
- 서울대학교 전자공학과 석사
- 위스콘신-메디슨 대학교 박사

주요이력
- **前** 삼성전자 메모리사업부 부사장
- **前** 바이오기업 대표이사