ISSN: 2181-3906 2026

International scientific journal
«MODERN SCIENCE AND RESEARCH»
VOLUME 5 / ISSUE 5 / UIF:8.2 / MODERNSCIENCE.UZ

SUN'IY INTELLEKT TIZIMLARIDA HISOBLASH SAMARADORLIGINI OSHIRISH UCHUN GPU VA NPU ARXITEKTURALARIDAN FOYDALANISH

Xatamov Xasanboy Sayidjon o'g'li
Soliyev Maqsudjon Murodjon o'g'li
Farg'ona Davlat texnika universiteti
Axborot texnologiyalari va telekommunikatsiya fakulteti
Suniy Intelekt yo'nalishi 3-bosqich 72-23SI-guruh talabasi
Sobirov Muzaffarjon Mirzaolimovich
(FDTU) Kompyuter muhandisligi va sun'iy intellekt kafedrasi (katta o'qituvchi) (ilmiy rahbar)

https://doi.org/10.5281/zenodo.20229302

Annotatsiya. *Ushbu maqolada sun'iy intellekt tizimlarida hisoblash samaradorligini oshirishda GPU va NPU arxitekturalarining o'rni ilmiy jihatdan tahlil qilinadi. Zamonaviy AI modellarining murakkablashuvi, katta hajmdagi ma'lumotlar bilan ishlashi va real vaqt rejimida qaror qabul qilish zarurati kompyuter arxitekturasiga bo'lgan talabni kuchaytirmoqda. An'anaviy CPU arxitekturasi umumiy maqsadli hisoblashlar uchun qulay bo'lsa-da, chuqur o'rganish modellarida uchraydigan massiv matritsali va tensor amallarni yuqori samaradorlikda bajarishda cheklovlarga ega. Shu sababli GPU arxitekturasi katta hajmli parallel hisoblashlar, modelni o'qitish va generativ AI tizimlarida keng qo'llanilmoqda. NPU arxitekturasi esa kam quvvat sarfi, past kechikish va qurilma ichida real vaqtli inferensiya bajarish imkoniyati bilan edge AI tizimlarida muhim o'rin egallaydi. Maqolada GPU va NPU arxitekturalarining afzalliklari, cheklovlari, xotira devori muammosi, energiya samaradorligi, kvantlash va gibrid GPU–NPU yondashuvining istiqbollari yoritiladi.*

Kalit so'zlar: *Sun'iy intellekt, GPU, NPU, kompyuter arxitekturasi, AI accelerator, parallel hisoblash, edge AI, xotira devori, energiya samaradorligi, neyron tarmoqlar, inferensiya, modelni optimallashtirish.*

Аннотация. *В данной статье научно анализируется роль архитектур GPU и NPU в повышении вычислительной эффективности систем искусственного интеллекта.*

Усложнение современных моделей ИИ, работа с большими объемами данных и необходимость принятия решений в режиме реального времени усиливают требования к компьютерной архитектуре. Традиционная архитектура CPU удобна для универсальных вычислений, однако имеет определенные ограничения при выполнении массивных матричных и тензорных операций, характерных для моделей глубокого обучения. В связи с этим архитектура GPU широко применяется для параллельных вычислений, обучения крупных моделей и работы генеративных систем искусственного интеллекта.

Архитектура NPU, в свою очередь, играет важную роль в системах edge AI благодаря низкому энергопотреблению, малой задержке и возможности выполнять инференс непосредственно на устройстве. В статье рассматриваются преимущества и ограничения GPU и NPU, проблема «стены памяти», энергоэффективность, квантование моделей, а также перспективы гибридного подхода GPU–NPU.

Ключевые слова: *Искусственный интеллект, GPU, NPU, компьютерная архитектура, AI-ускоритель, параллельные вычисления, edge AI, стена памяти, энергоэффективность, нейронные сети, инференс, оптимизация моделей.*

Abstract. *This article scientifically analyzes the role of GPU and NPU architectures in improving computational efficiency in artificial intelligence systems. The increasing complexity of modern AI models, their dependence on large-scale data processing, and the need for real-time decision-making have significantly increased the demand for advanced computer architectures. Although traditional CPU architecture is suitable for general-purpose computing, it has limitations in efficiently performing massive matrix and tensor operations commonly used in deep learning models. Therefore, GPU architecture is widely used for large-scale parallel computing, model training, and generative AI systems. NPU architecture, on the other hand, plays an important role in edge AI systems due to its low power consumption, low latency, and ability to perform real-time inference directly on devices. The article discusses the advantages and limitations of GPU and NPU architectures, the memory wall problem, energy efficiency, model quantization, and the prospects of a hybrid GPU–NPU approach.*

Keywords: *Artificial intelligence, GPU, NPU, computer architecture, AI accelerator, parallel computing, edge AI, memory wall, energy efficiency, neural networks, inference, model optimization.*

Kirish

So‘nggi yillarda sun’iy intellekt tizimlari oddiy dasturiy modullardan murakkab hisoblash platformalariga aylandi. Katta til modellari, tibbiy diagnostika, avtonom transport, aqlli kamera, robototexnika va mobil qurilmalardagi AI funksiyalari katta hajmdagi ma’lumotlarni real vaqtga yaqin tezlikda qayta ishlashni talab qiladi. Bunday vazifalarda asosiy hisoblash operatsiyalari ko‘pincha matritsa ko‘paytirish, konvolyutsiya, vektor amallari va tensor hisoblashlardan iborat bo‘ladi.

An’anaviy CPU arxitekturasi murakkab boshqaruv, ketma-ket buyruqlar va umumiy maqsadli hisoblashlar uchun ishlab chiqilgan. Lekin sun’iy intellekt modellarida minglab yoki millionlab o‘xshash matematik amallar bir vaqtning o‘zida bajariladi. Bunday holatda GPU arxitekturasi o‘zining ko‘p sonli yadrolari va parallel hisoblash imkoniyati bilan samarali yechim bo‘la oladi. NVIDIA H100 kabi zamonaviy GPUlarda Tensor Core bloklari FP8 formatida AI training va inference vazifalarini tezlashtirishga mo‘ljallangan; NVIDIA Hopper arxitekturasi H100’da FP8 Tensor Corelarni qo‘shganini bildiradi .

Biroq GPUlar yuqori hisoblash quvvatiga ega bo‘lsa-da, ular katta energiya sarfi va kuchli sovitish tizimini talab qiladi. Shu sababli mobil qurilmalar, IoT tizimlari va edge AI muhitlarida NPU arxitekturasi dolzarb bo‘lib bormoqda. Masalan, Qualcomm Hexagon NPU qurilma ichida AI inference bajarish uchun maxsus ishlab chiqilganini ta’kidlaydi . Snapdragon 8 Elite platformasida Qualcomm Hexagon NPU 45% AI unumdorlik o‘sishi va 45% yaxshi performance-per-watt ko‘rsatkichini bergani ko‘rsatilgan .

Shu jihatdan maqolaning dolzarbligi shundaki, sun’iy intellekt tizimlarining rivojlanishi nafaqat algoritmlarga, balki ularni bajaradigan kompyuter arxitekturasiga ham bevosita bog‘liq.

ISSN: 2181-3906 2026

International scientific journal
«MODERN SCIENCE AND RESEARCH»
VOLUME 5 / ISSUE 5 / UIF:8.2 / MODERNSCIENCE.UZ

AI tizimlari uchun asosiy savol: qaysi vazifa GPUda, qaysi vazifa NPUda, qaysi holatda esa gibrid arxitekturada bajarilishi kerak?

Tadqiqot maqsadi va vazifalari

Ushbu maqolaning asosiy maqsadi — sun’iy intellekt tizimlarida hisoblash samaradorligini oshirishda GPU va NPU arxitekturalarining imkoniyatlarini ilmiy jihatdan tahlil qilish.

Tadqiqot vazifalari quyidagilardan iborat:

GPU va NPU arxitekturalarining ishlash tamoyillarini tahlil qilish.

AI vazifalarida parallel hisoblash va tensor amallarining ahamiyatini yoritish.

GPU va NPUning afzalliklari hamda cheklovlarini solishtirish.

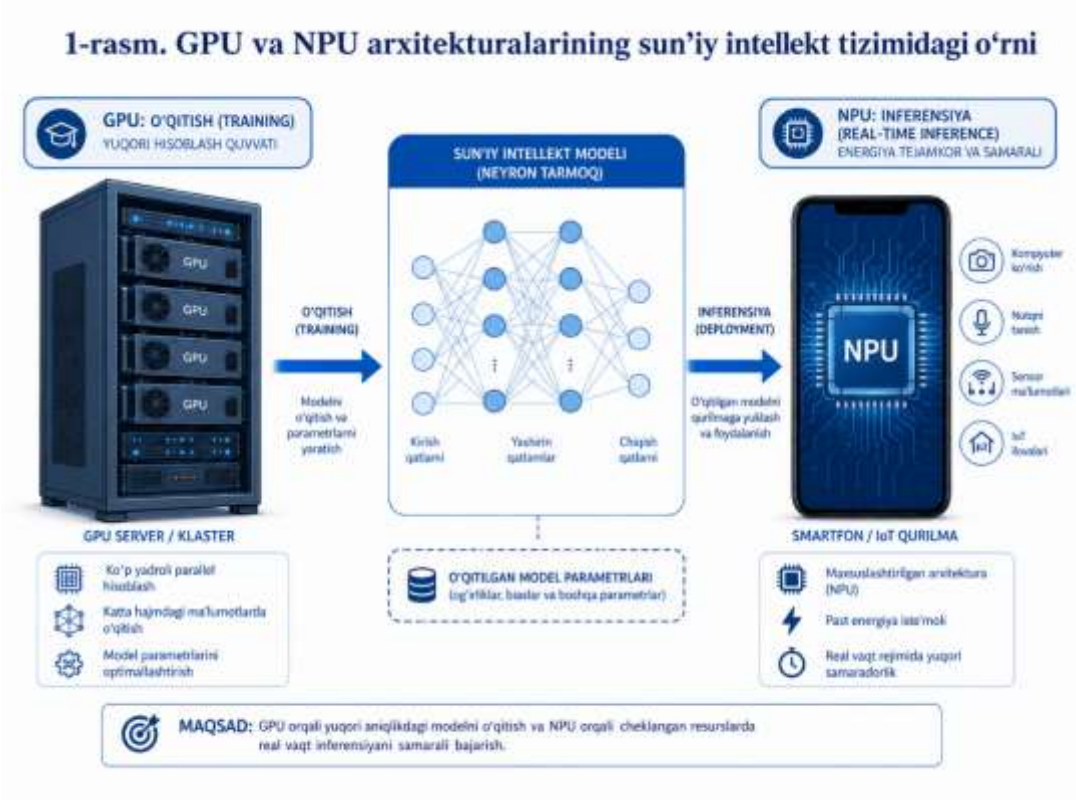
Hisoblash samaradorligiga ta’sir qiluvchi xotira, energiya va dasturiy omillarni aniqlash.

Hozirgacha to‘liq yechilmagan muammolarni ko‘rsatish.

AI tizimlari uchun gibrid GPU–NPU arxitektura modelini taklif qilish.

1-rasm

Nomi: GPU va NPU arxitekturalarining sun’iy intellekt tizimidagi o‘rni



GPU va NPU arxitekturalarining sun’iy intellekt tizimidagi o‘rni.

Tavsifi:

*Ushbu rasmda sun’iy intellekt tizimida **GPU** va **NPU** ning asosiy vazifalari ko‘rsatiladi.*

*Chap tomonda GPU server joylashgan bo‘lib, u **modelni o‘qitish (training)** jarayonini bajaradi. Markazda **sun’iy intellekt modeli (neyron tarmoq)** berilgan. O‘ng tomonda esa **NPU bilan jihozlangan smartfon yoki IoT qurilma** tasvirlanib, unda **real vaqtli inferensiya** bajarilishi ko‘rsatilgan. Rasm maqolaning asosiy g‘oyasini ochib beradi.*

Metodlar

Maqola IMRAD talablariga mos ravishda yozildi. Tadqiqotda quyidagi metodlardan foydalanildi:

Tahliliy metod: GPU, NPU, TPU va CPU arxitekturalari nazariy jihatdan solishtirildi.

Qiyosiy metod: AI workloadlar — training, inference, edge AI va real vaqtli qayta ishlash bo‘yicha arxitekturalarning imkoniyatlari baholandi.

Manbaviy tahlil: NVIDIA, Google Cloud, Qualcomm, MLCommons va ilmiy maqolalardagi ma’lumotlar asosida dolzarb texnik ko‘rsatkichlar o‘rganildi.

Sistemali yondashuv: Muammo faqat chip tezligida emas, balki xotira iyerarxiyasi, energiya samaradorligi, dasturiy ekotizim va model optimallashtirish bilan bog‘liq holda tahlil qilindi.

MLCommons MLPerf benchmarklari turli AI tizimlarining training va inference unumdorligini belgilangan sharoitlarda baholash uchun ishlab chiqilgan arxitektura-neytral testlar sifatida tavsiflanadi. Shu sababli AI arxitekturalarini solishtirishda faqat nazariy FLOPS emas, balki real workload natijalari ham muhim hisoblanadi.

Nazariy asos

CPU arxitekturasi

CPU — umumiy maqsadli protsessor bo‘lib, operatsion tizim, ilovalar, boshqaruv algoritmlari va ketma-ket buyruqlarni bajarishda qulay. CPUning asosiy kuchli tomoni — moslashuvchanlik. Lekin chuqur o‘rganish modellaridagi millionlab parallel matritsa amallarida CPU samaradorligi GPU yoki NPUga nisbatan past bo‘ladi.

GPU arxitekturasi

GPU dastlab grafik hisoblashlar uchun yaratilgan bo‘lsa-da, bugungi kunda sun’iy intellekt training va inference vazifalarida asosiy tezlatkichlardan biriga aylandi. GPU minglab parallel yadrolar orqali bir xil turdagi matematik amallarni bir vaqtda bajaradi.

GPUning AI uchun muhim imkoniyatlari:

massiv parallel hisoblash;

Tensor Core bloklari;

yuqori xotira o‘tkazuvchanligi;

FP16, BF16, FP8 kabi past aniqlikdagi formatlarni qo‘llash;

CUDA, cuDNN, TensorRT kabi dasturiy ekotizimlar.

NVIDIA H100 sahifasida H100 Tensor Core GPU HPC va AI vazifalari uchun yuqori unumdorlik berishi, jumladan TF32 formatida bir petaflopgacha matrix-multiply throughput taqdim etishi ko‘rsatilgan . Blackwell arxitekturasida esa NVIDIA FP4 AI hisoblash formatini qo‘llab, model sig‘imi va unumdorligini oshirishga urg‘u beradi .

NPU arxitekturasi

NPU — neural processing unit, ya’ni neyron tarmoqlarni bajarish uchun maxsuslashtirilgan protsessor. GPU ko‘proq umumiy parallel hisoblashlarda kuchli bo‘lsa, NPU aynan AI inference vazifalarini kam quvvat bilan bajarishga yo‘naltiriladi.

3-rasm

Nomi: Edge AI qurilmasida NPU ishlash jarayoni

Tavsifi:

ISSN: 2181-3906 2026

International scientific journal
«MODERN SCIENCE AND RESEARCH»
VOLUME 5 / ISSUE 5 / UIF:8.2 / MODERNSCIENCE.UZ

*Bu rasmda **kamera**, **sensorlar** va boshqa kirish manbalaridan keladigan ma'lumotlar **NPU** ga uzatilishi, NPU ichida **neyron tarmoq inferensiyasi** bajarilishi va natijada **real vaqtli qaror yoki chiqish** hosil bo'lishi ko'rsatiladi. Rasm Edge AI qurilmalarida NPU qanday ishlashini ilmiy jihatdan tushuntiradi. Shuningdek, uning afzalliklari — **past quvvat sarfi**, **tezkor javob**, **internetga bog'liqlikning kamligi** ham ifodalangan.*

3-rasm. Edge AI qurilmasida NPU ishlash jarayoni.

NPUning asosiy xususiyatlari:

past energiya sarfi;

mobil va edge qurilmalarda ishlash;

qurilma ichida AI inference bajarish;

maxsus matritsa va tensor bloklari;

real vaqtli ovoz, kamera va sensor ma'lumotlarini qayta ishlash.

Masalan, Qualcomm Hexagon NPU on-device AI inferencing uchun maxsus ishlab chiqilganini ko'rsatadi . Bu mobil telefon, aqlli kamera, dron va IoT qurilmalarda bulutga murojaat qilmasdan AI funksiyalarini bajarish imkonini beradi.

TPU va boshqa AI acceleratorlar

TPU — Google tomonidan ishlab chiqilgan Tensor Processing Unit bo'lib, asosan tensor va matritsa hisoblashlarga mo'ljallangan. Google TPU v5p avlodida oldingi TPU v4ga nisbatan FLOPS ko'rsatkichi 2 martadan ortiq, HBM xotirasi esa 3 marta ko'proq ekanini bildirgan.

Trillium TPU esa Google ma'lumotiga ko'ra TPU v5ega nisbatan 4,7 baravar yuqori peak compute performance va 67% yuqori energiya samaradorligini beradi .

Natijalar

GPU va NPU arxitekturalarining taqqoslanishi

MezON	GPU	NPU
Asosiy vazifa	Parallel hisoblash, AI	Kam quvvatli AI inference

ISSN: 2181-3906 2026

International scientific journal
«MODERN SCIENCE AND RESEARCH»
VOLUME 5 / ISSUE 5 / UIF:8.2 / MODERNSCIENCE.UZ

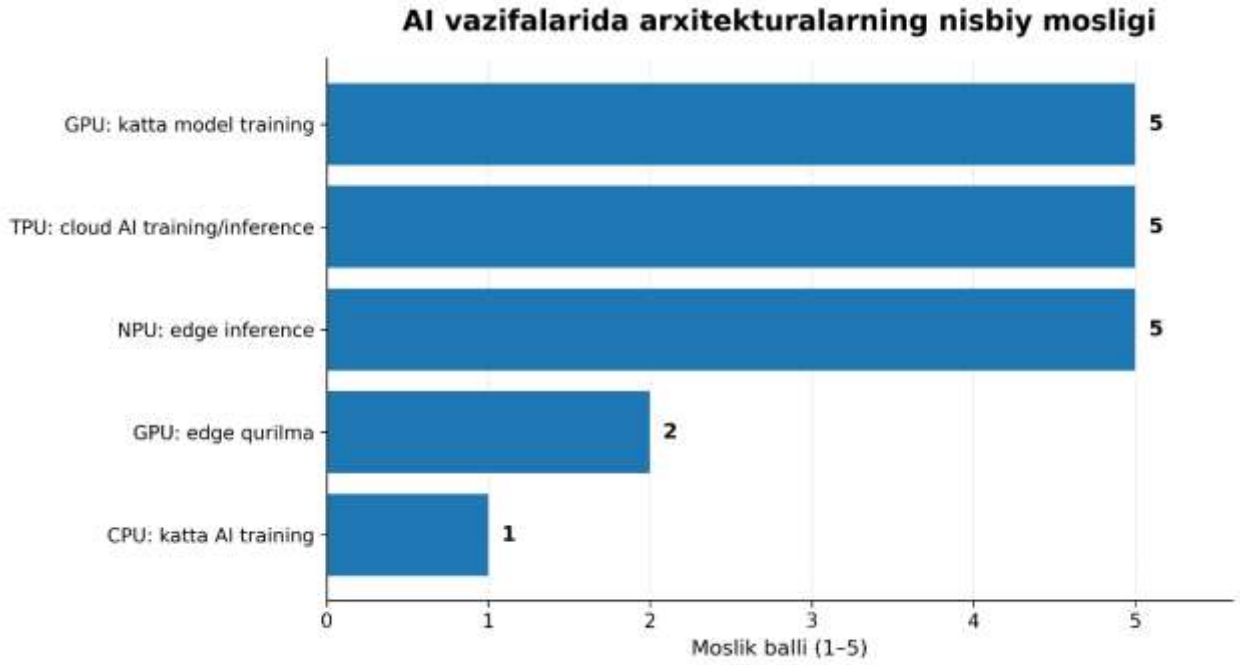
	training, katta inference	
Kuchli tomoni	Katta model training, yuqori throughput	Energiya tejamkorlik, edge AI
Ishlatilish joyi	Data center, server, workstation	Smartfon, IoT, kamera, avtomobil
Model turi	CNN, Transformer, LLM, multimodal AI	Optimallashtirilgan CNN, kichik LLM, ovoz/kamera AI
Energiya sarfi	Yuqori	Nisbatan past
Dasturiy ekotizim	CUDA, TensorRT, PyTorch, TensorFlow	Vendor SDK, Android NNAPI, Core ML, Qualcomm AI Stack
Cheklovi	Issiqlik, narx, energiya sarfi	Moslashuvchanlik pastroq, model sig‘imi cheklangan
Eng mos holat	Training va katta hajmli inference	Mobil va real vaqtli inference

Arxitekturalarning vazifa bo‘yicha nisbiy mosligi

Quyidagi diagramma konseptual baholash sifatida berildi. 5 ball — eng yuqori moslik, 1 ball — past moslik.

AI vazifalarida arxitekturalarning nisbiy mosligi

1–5 ballik metodik baholash: training, data-center inference va edge inference bo‘yicha umumiy moslik.



1-diagramma. AI vazifalarida arxitekturalarning nisbiy mosligi.

Asosiy ilmiy natijalar

Tahlil shuni ko‘rsatadiki, GPU sun’iy intellekt tizimlarida yuqori tezlikdagi parallel hisoblashlar uchun samarali. Ayniqsa, Transformer arxitekturasi, CNN, katta til modellari va generativ AI tizimlarida GPUlar yuqori throughput beradi. Biroq GPUlar ko‘p energiya talab qiladi va ularni kichik qurilmalarda ishlatish murakkab.

NPU esa AI inference vazifalarida, ayniqsa real vaqtli, kam quvvatli va qurilma ichidagi hisoblashlarda samaralidir. Mobil qurilmalar, aqlli kameralar, sensor tizimlar va IoT qurilmalarida NPUdan foydalanish kechikishni kamaytiradi, internetga bog‘liqlikni pasaytiradi va maxfiylikni yaxshilaydi.

TPU va boshqa maxsus AI acceleratorlar esa cloud muhitda katta model training va inference vazifalarini optimallashtirishga xizmat qiladi. Google Trillium TPU avlodida energiya samaradorligi va peak compute performance keskin oshgani ko‘rsatilgan .

Ko‘rsatkich	**GPU**	**NPU**
Hisoblash turi	Keng ko‘lamli parallel hisoblash	Neyron tarmoqqa maxsus inference
Afzalligi	Training va katta modellar uchun kuchli	Kam energiya, real vaqtli ishlash
Kamchiligi	Energiya sarfi yuqori	Model moslashuvchanligi pastroq
Qo‘llanish sohasi	Data center, ilmiy hisoblash, generativ AI	Smartfon, IoT, robot, kamera
Model hajmi	Katta va murakkab modellar	Optimallashtirilgan kichik/moderate modellar
Energiya samaradorligi	O‘rtacha yoki yuqori sarfli	Yuqori tejamkor
Kelajakdagi roli	AI factory va cloud training	Edge AI va shaxsiy AI qurilmalar

Muhokama

Dolzarb muammo: xotira devori

2-rasm

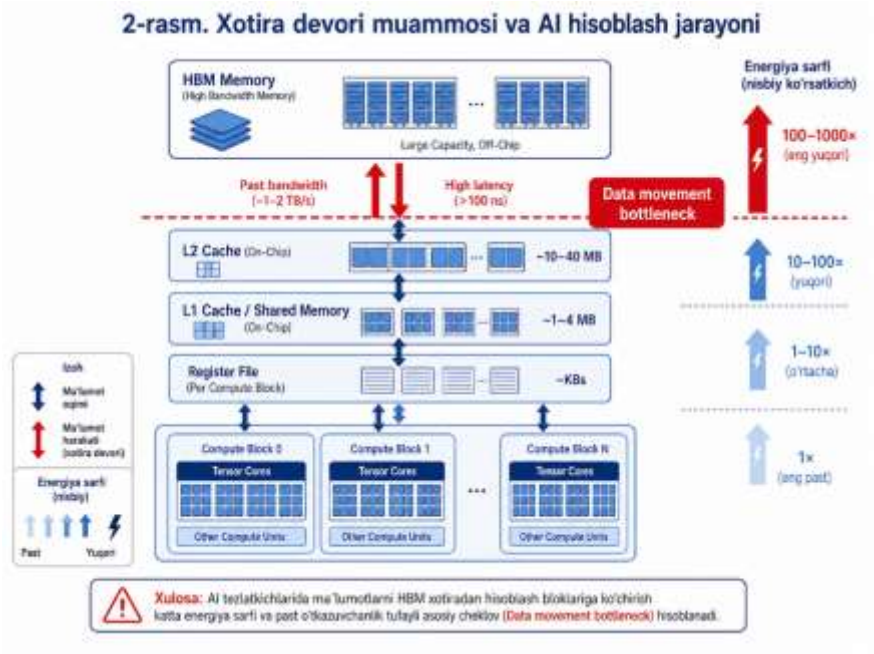
Nomi: *Xotira devori muammosi va AI hisoblash jarayoni*

Tavsifi:

Mazkur rasmda AI tezlatkichlarida ma’lumotlarning ***HBM xotira****,* ***L2 kesh****,* ***L1 kesh / shared memory****,* ***register file*** *va* ***tensor yadro****lar orasida harakati ko‘rsatiladi. Aynan ma’lumotlarni xotiradan hisoblash bloklariga ko‘chirishdagi sekinlashuv* ***“data movement bottleneck”*** *sifatida ajratib ko‘rsatilgan. Rasm AI hisoblash tizimlarida unumdorlik va energiya sarfiga xotira devori qanday ta’sir qilishini tushuntiradi.*

ISSN: 2181-3906 2026

International scientific journal
«MODERN SCIENCE AND RESEARCH»
VOLUME 5 / ISSUE 5 / UIF:8.2 / MODERNSCIENCE.UZ



Xotira devori muammosi va AI hisoblash jarayoni.

AI tizimlarida eng katta muammolardan biri — memory wall, ya’ni xotira devori muammosidir. Bu muammo shundan iboratki, protsessor yoki tezlatkichning hisoblash bloklari juda tez ishlashi mumkin, ammo ma’lumotlarni xotiradan olib kelish va qayta yozish jarayoni umumiy tezlikni cheklaydi.

“AI and Memory Wall” nomli ilmiy ishda katta til modellarida xotira bandwidthi decoder modellar uchun asosiy bottleneckga aylanishi mumkinligi ta’kidlanadi. SC23 materiallarida esa neyron tarmoq tezlatkichlarida og‘irliklar va aktivatsiyalarni xotiradan olib kelish jarayoni eng katta unumdorlik va energiya cheklovi bo‘lishi, an’anaviy von Neumann arxitekturalarida energiyaning katta qismi ma’lumot harakatiga ketishi qayd etiladi .

Demak, kelajakdagi AI arxitekturalarida faqat ko‘proq yadro qo‘shish yetarli emas. Asosiy yechimlar quyidagilar bo‘lishi kerak:

xotiraga yaqin hisoblash;
in-memory computing;
model kvantlash;
sparse computation;
xotira iyerarxiyasini optimallashtirish;
ma’lumotlarni qayta ishlatish strategiyalari.
Energiya samaradorligi muammosi

AI tizimlarining keng qo‘llanilishi data centerlarda elektr energiyasi sarfini oshirmoqda. GPUlar juda kuchli bo‘lsa-da, katta modellarni o‘qitish va ishga tushirishda sovitish, elektr ta’minoti va infratuzilma xarajatlari ortadi. Shu sababli performance-per-watt ko‘rsatkichi AI arxitekturasining eng muhim mezonlaridan biriga aylanmoqda.

NPUlar aynan shu muammoga javob sifatida paydo bo‘lgan. Ular kam quvvat sarflab, qurilma ichida AI inference bajaradi.

Google Trillium TPU ham TPU v5ega nisbatan 67% yuqori energiya samaradorligi bilan taqdim etilgan . Qualcomm esa Snapdragon 8 Elite platformasida Hexagon NPUning 45% yaxshi performance-per-watt ko‘rsatkichini bildirgan .

Kvantlash va aniqlik muammosi

AI modellarini tezlashtirishda FP32 o‘rniga FP16, BF16, FP8, INT8 yoki FP4 kabi past aniqlikdagi formatlardan foydalaniladi. Bu hisoblash tezligini oshiradi va xotira sarfini kamaytiradi. NVIDIA Blackwell Transformer Engine FP4 AI formatini qo‘llab, model sig‘imi va performance samaradorligini oshirishga urg‘u beradi .

Biroq past aniqlikdagi formatlardan foydalanish har doim ham muammosiz emas. Ayrim modellar, ayniqsa tibbiyot, xavfsizlik, huquqiy tahlil yoki moliyaviy qarorlar kabi yuqori aniqlik talab qiladigan sohalarda kvantlash natijasida xatolik ehtimoli oshishi mumkin. Shu sababli yechilmagan ilmiy muammo quyidagicha: qanday qilib modelni maksimal darajada siqish va tezlashtirish mumkin, lekin aniqlikni sezilarli kamaytirmaslik kerak?

Dasturiy ekotizim muammosi

GPUlarning ustunligi faqat apparat kuchida emas, balki dasturiy ekotizimidadir. CUDA, cuDNN, TensorRT, PyTorch va TensorFlow kabi vositalar GPUlardan foydalanishni qulaylashtiradi. NPUlarda esa vaziyat murakkabroq: har bir ishlab chiqaruvchi o‘z SDKsi, kompilyatori va optimallashtirish vositalarini taklif qiladi. Bu esa modelni bir qurilmadan boshqasiga ko‘chirishda muammo tug‘diradi.

Bu yo‘nalishda ochiq standartlar, universal kompilyatorlar, ONNX, TVM, MLIR, Android NNAPI va vendorlararo moslik muhim ahamiyatga ega.

Taklif etilayotgan yechim: gibrid GPU–NPU yondashuvi

Maqola tahliliga ko‘ra, AI tizimlarida bitta arxitektura barcha vazifalar uchun ideal yechim bo‘la olmaydi. Eng maqbul yondashuv — gibrid GPU–NPU arxitekturasi.

Ma’lumotlar
↓
Oldindan qayta ishlash
↓
Modelni o‘qitish / fine-tuning
↓
GPU yoki TPU
↓
Modelni optimallashtirish
↓
Kvantlash, pruning, distillation
↓
Edge qurilmaga joylash
↓
NPU orqali inference
↓
Real vaqtli qaror

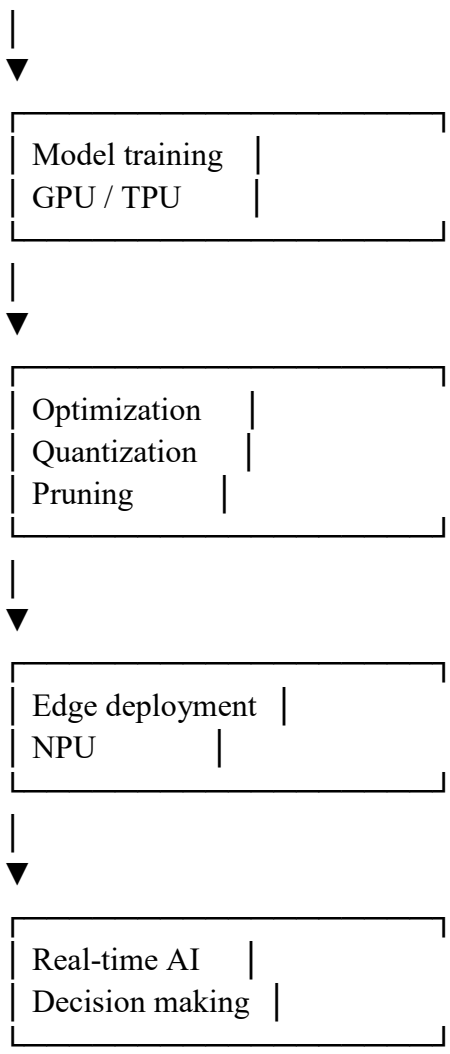
Modelning afzalliklari

Bosqich	**Tavsiya etiladigan arxitektura**	**Sabab**
Katta modelni o‘qitish	GPU yoki TPU	Yuqori parallel hisoblash quvvati
Fine-tuning	GPU	Moslashuvchanlik va framework qo‘llab-quvvatlashi
Modelni siqish	GPU/CPU	Kvantlash va optimallashtirish jarayoni
Mobil inference	NPU	Kam quvvat, past kechikish
Real vaqtli sensor AI	NPU	Qurilma ichida tez javob
Cloud inference	GPU/TPU	Katta foydalanuvchi yuklamasi

Diagramma nomi

1-diagramma. *Sun’iy intellekt tizimlarida GPU va NPU asosidagi gibrid hisoblash modeli*

Diagramma ko‘rinishi



ISSN: 2181-3906 2026

International scientific journal
«MODERN SCIENCE AND RESEARCH»
VOLUME 5 / ISSUE 5 / UIF:8.2 / MODERNSCIENCE.UZ

Hozirgacha to‘liq yechilmagan muammolar

Ushbu mavzuda quyidagi ilmiy va amaliy muammolar hanuz dolzarb hisoblanadi:

Xotira devori muammosi: Hisoblash bloklari tezlashmoqda, ammo xotiradan ma’lumot uzatish tezligi va energiya sarfi umumiy samaradorlikni cheklamoqda.

Energiya va sovitish muammosi: Katta GPU klasterlari yuqori elektr quvvati va murakkab sovitish tizimini talab qiladi.

Kvantlashdagi aniqlik yo‘qotilishi: FP8, INT8 yoki FP4 formatlari tezlikni oshiradi, lekin ayrim modellar aniqligiga salbiy ta’sir qilishi mumkin.

NPUlar orasidagi standartlashuv muammosi: Har bir ishlab chiqaruvchi o‘z NPU ekotizimini yaratmoqda, bu esa model portativligini qiyinlashtiradi.

Edge AI xavfsizligi: Qurilma ichida AI ishlashi maxfiylikni oshiradi, lekin modelni himoya qilish, adversarial hujumlarga bardoshlilik va yangilash muammolari mavjud.

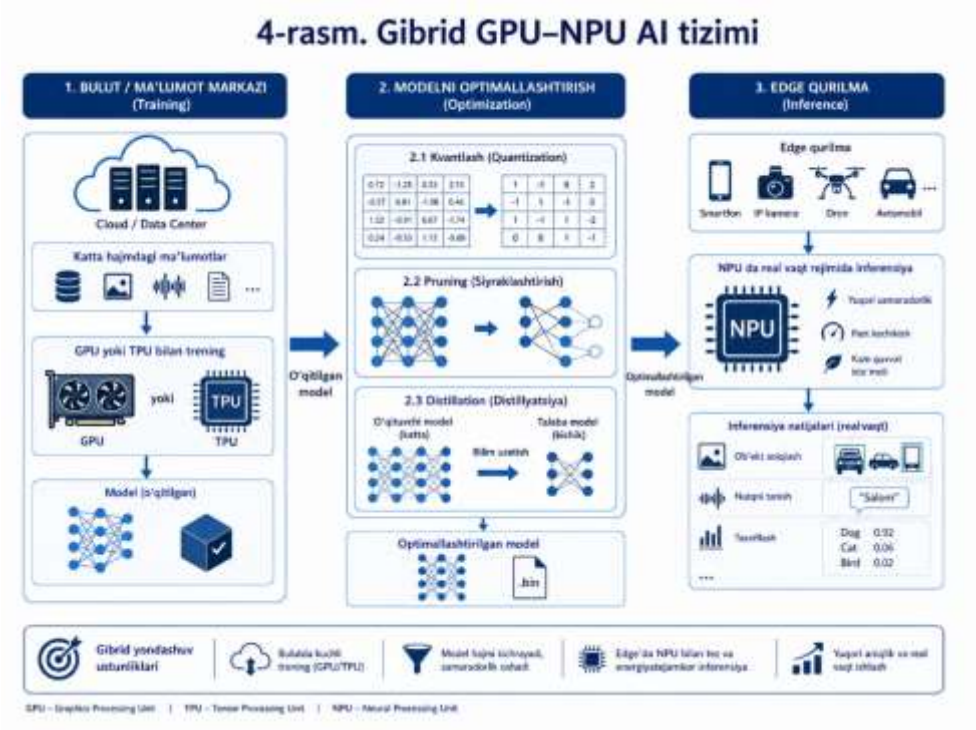
Real benchmark muammosi: Nazariy TOPS yoki FLOPS ko‘rsatkichi har doim real AI ilova tezligini to‘liq aks ettirmaydi. Shu sababli MLPerf kabi amaliy benchmarklar muhim hisoblanadi.

4-rasm

Nomi: Gibrid GPU–NPU AI tizimi

Tavsifi:

*Mazkur rasmda **gibrid GPU–NPU arxitektura modeli** ko‘rsatiladi. Avval model **bulut yoki ma’lumot markazida GPU/TPU yordamida o‘qitiladi**, keyin **kvantlash, pruning, distillation** kabi optimallashtirish bosqichlaridan o‘tkaziladi. Shundan so‘ng optimallashtirilgan model **edge qurilmaga uzatiladi** va **NPU orqali real vaqtli inferensiya** bajariladi. Bu rasm maqoladagi taklif qilinayotgan yechimni ifodalovchi asosiy rasm hisoblanadi.*



4-rasm. Gibrid GPU–NPU AI tizimi.

Xulosa

Sun’iy intellekt tizimlarida hisoblash samaradorligini oshirish zamonaviy kompyuter arxitekturasining eng dolzarb yo‘nalishlaridan biridir. Tahlillar shuni ko‘rsatadiki, GPU arxitekturasi katta hajmli AI modellarini o‘qitish va yuqori throughput talab qiladigan inference vazifalarida samarali hisoblanadi. GPUning parallel hisoblash imkoniyati, Tensor Core bloklari va rivojlangan dasturiy ekotizimi uni generativ AI, katta til modellari va data center tizimlarida muhim platformaga aylantiradi.

NPU esa edge AI, mobil qurilmalar, aqlli sensorlar va real vaqtli inference vazifalarida dolzarbdir. Uning asosiy ustunligi — kam quvvat sarfi, past kechikish va qurilma ichida ishlash imkoniyatidir. Shu bilan birga, NPUlar hali to‘liq standartlashmagan, ularning dasturiy ekotizimi GPUlarga nisbatan cheklanganroq.

Maqolaning asosiy ilmiy xulosasi shuki: kelajakdagi AI tizimlari uchun eng maqbul yechim — GPU, TPU va NPU imkoniyatlarini birlashtirgan gibrid arxitekturadir. Bunda katta model cloud yoki data centerda GPU/TPU yordamida o‘qitiladi, so‘ng optimallashtirilgan model edge qurilmaga joylashtirilib, NPU orqali real vaqtli inference bajariladi. Bunday yondashuv hisoblash tezligi, energiya samaradorligi, maxfiylik va amaliy qo‘llanish imkoniyatlarini muvozanatlashtiradi.

FOYDALANILGAN ADABIYOTLAR:

1. NVIDIA. (2024). NVIDIA H100 Tensor Core GPU. NVIDIA rasmiy ma’lumotlar sahifasi. H100 GPU sun’iy intellekt va yuqori unumdor hisoblash vazifalari uchun mo‘ljallangan bo‘lib, Tensor Core va FP8 imkoniyatlarini qo‘llab-quvvatlaydi.
2. NVIDIA. (2024). NVIDIA Hopper GPU Architecture. NVIDIA rasmiy texnik ma’lumotlari. Hopper arxitekturasi Transformer Engine va FP8/FP16 aralash aniqlikdagi hisoblashlar orqali AI modellarini tezlashtirishga qaratilgan.
3. NVIDIA Developer. (2022). NVIDIA Hopper Architecture In-Depth. NVIDIA Developer Blog. Maqolada H100 GPUning to‘rtinchi avlod Tensor Corelari, FP8, FP16, BF16, TF32 va INT8 hisoblash formatlari haqida ma’lumot berilgan.
4. NVIDIA. (2025). NVIDIA Blackwell Architecture. NVIDIA rasmiy sahifasi. Blackwell arxitekturasi katta til modellari va Mixture-of-Experts modellarini o‘qitish hamda inferensiya qilish uchun yangi Tensor Core va Transformer Engine imkoniyatlarini taqdim etadi.
5. NVIDIA Documentation. (2025). Using FP8 and FP4 with Transformer Engine. NVIDIA hujjatlari. Manbada Blackwell arxitekturasida FP8 va FP4 kabi past aniqlikdagi formatlardan foydalanish imkoniyatlari yoritilgan.
6. Google Cloud. (2024). Trillium TPU is GA. Google Cloud Blog. Trillium TPU oldingi avlodga nisbatan yuqori training performance, inference throughput va energiya samaradorligini taqdim etishi ko‘rsatilgan.
7. Google Cloud. (2025). Tensor Processing Units (TPUs). Google Cloud rasmiy sahifasi. TPUlar sun’iy intellekt modellarini training va inference qilish uchun mo‘ljallangan maxsus tezlatkichlar sifatida tavsiflanadi.

8. Google Cloud Documentation. (2025). TPU v5p Architecture. Google Cloud hujjatlari. TPU v5p arxitekturasi, TensorCore, Matrix Multiply Unit va pod konfiguratsiyalari haqida texnik ma’lumot beradi.
9. Qualcomm. (2024). Snapdragon 8 Elite Mobile Platform. Qualcomm rasmiy sahifasi. Snapdragon 8 Elite platformasida Hexagon NPUning 45% yaxshilangan AI unumdorligi va 45% yuqori performance-per-watt ko‘rsatkichi keltirilgan.
10. Qualcomm. (2024). Snapdragon 8 Elite Platform Product Brief. Qualcomm texnik hujjati. Unda Hexagon NPU, on-device generative AI, multimodal qo‘llab-quvvatlash va energiya samaradorligi haqida ma’lumot berilgan.
11. MLCommons. (2025). MLPerf Training Benchmark. MLCommons rasmiy benchmark sahifasi. MLPerf Training benchmarklari tizimlarning AI modellarini belgilangan sifat ko‘rsatkichigacha qanchalik tez o‘qitishini baholaydi.
12. MLCommons. (2025). MLPerf Inference: Datacenter Benchmark. MLCommons rasmiy sahifasi. Ushbu benchmark trained model yordamida tizimlarning kiritilgan ma’lumotlardan natija hosil qilish tezligini baholaydi.
13. NVIDIA. (2025). MLPerf AI Benchmarks. NVIDIA Data Center Resources. MLPerf benchmarklari hardware, software va AI xizmatlarining training hamda inference samaradorligini xolis baholash uchun ishlatiladi.
14. Gholami, A., Yao, Z., Kim, S., Hooper, C., Mahoney, M. W., & Keutzer, K. (2024). AI and Memory Wall. arXiv preprint. Ushbu tadqiqotda Transformer modellarida xotira bandwidthi asosiy bottleneckga aylanishi mumkinligi ilmiy asosda tahlil qilingan.
15. NVIDIA Developer. (2025). Introducing NVFP4 for Efficient and Accurate Low-Precision Inference. NVIDIA Developer Blog. Manbada Blackwell Tensor Core arxitekturasida NVFP4 formatidan samarali past aniqlikdagi inferensiya uchun foydalanish yoritilgan.