ChatGPT

# Alternatywne podejście sprzętowe z wykorzystaniem superpozycji sygnałów

## Wprowadzenie

Projekt **TITANS/PIAST-Q** stawia ogromne wymagania obliczeniowe – w analizie stwierdzono, że *„Faza 2 wymaga klastrów H100; koszt CAPEX/OPEX wysoki”*. Klastry GPU NVIDIA H100 oferują co prawda ogromną moc (~495 TFLOPS FP16, ~1000 TFLOPS FP8 na jeden akcelerator [1]), lecz są kosztowne (dziesiątki tysięcy dolarów za jednostkę) i pobierają setki watów mocy. Ponadto, modele kognitywne TITANS składają się z wielu modułów (M1–M5 a potencjalnie aż M1–M8) działających w pętli percepcja–pamięć–rozumowanie–agencja, co wymaga **minimalizacji opóźnień** w przepływie danych między modułami. Zatem poszukujemy alternatywnych rozwiązań sprzętowych opartych na **superpozycji sygnałów**, które mogłyby zapewnić *niższą latencję* i *wyższą przepustowość* przy niższym koszcie energetycznym, wspierając jednocześnie **wielomodelowe arbitraże LLM** (tj. równoczesne działanie wielu modeli i ich koordynację). Poniżej dokonujemy przeglądu takich technologii – od fotonicznych układów optycznych, przez neuromorficzne i analogowe układy krzemowe, po kwantowe procesory QPU – oceniając ich potencjał względem wymagań TITANS/PIAST-Q.

## Obliczenia fotoniczne (optyczne) na zasadzie interferencji

Układy fotoniczne wykorzystują **interferencję światła (superpozycję fal optycznych)** do wykonywania operacji matematycznych. W praktyce implementuje się to np. sieciami interferometrów Mach-Zehndera, które realizują mnożenia macierzowe dzięki nakładaniu się sygnałów optycznych. **Zalety** takiego podejścia to *bardzo wysokie pasmo i równoległość* oraz *niska latencja* – sygnał propaguje z prędkością światła, a wiele operacji może być wykonywanych jednocześnie (różnymi długościami fali) [2]. Fotony nie generują także strat mocy na rezystancjach jak prąd elektryczny, co daje potencjał *wysokiej efektywności energetycznej* [2].

**Przykładem** jest firma **Lightmatter**, która zademonstrowała w 2025 r. pierwszy uniwersalny procesor fotoniczny zdolny do wykonywania zaawansowanych sieci neuronowych (ResNet, BERT, RL Atari) bez modyfikacji modeli [3]. Układ ten, integrujący fotoniczne rdzenie tensorowe z elektroniką sterującą w jednym pakiecie, osiąga **65,5 bln operacji na sekundę (ABFP16)** przy poborze tylko **~78 W energii elektrycznej + 1,6 W optycznej** [4]. Osiągnięto przy tym **dokładność zbliżoną do 32-bitowej** na wyjściu modeli, co historycznie było barierą dla optycznych akceleratorów [3] [5]. Tak wysoka wydajność (65 TOPS) przy niskiej mocy oznacza efektywność ~0,8 TOPS/W, co już teraz porównywalne jest z GPU (H100 ~0,7 TOPS/W dla FP16). Latencja inferencji może ulec radykalnemu skróceniu – mnożenie macierzy realizowane jest praktycznie **w jednym przebiegu światła przez układ** (nanosekundy), podczas gdy w GPU operacja ta wymaga setek cykli zegara. Co więcej, fotonika umożliwia *„color-enabled parallelism”* – wiele długości fali niesie równolegle różne dane [2] – co potencjalnie pozwala wykonywać *kilka strumieni obliczeń jednocześnie* na jednym rdzeniu. Tę cechę można wykorzystać do równoczesnego wykonywania wielu modułów (np. wielomodelowych ekspertów LLM) i następnie **sprzętowego arbitrażu** ich wyników w czasie niemal rzeczywistym.

**Wyzwania:** Dzisiejsze fotoniczne układy wciąż wymagają konwersji opto-elektronicznej (dane wejściowe muszą być zakodowane w modulatory, a wyjścia odczytane fotodetektorami), co wprowadza pewne

opóźnienie i koszt energetyczny. Ponadto programowanie takich procesorów wymaga nowych kompilatorów i metod optymalizacji – układ Lightmatter korzysta ze specjalnych formatu liczby (Adaptive Block Floating Point 16) i musi skalibrować interferometry. Niemniej jednak, najnowsze wyniki pokazują, że **fotoniczne akceleratory mogą już dziś bezpośrednio uruchamiać transformery i konwolucyjne sieci** z dokładnością porównywalną do GPU [3] [5]. Firma Lightmatter rozwija także platformę **Passage**™ – fotoniczną magistralę międzyukładową o przepustowości do 32–64 Tb/s [6] – co pozwoli skalować system przez łączenie wielu fotonicznych chipów z CPU/GPU. Dla TITANS oznacza to możliwość zbudowania hybrydowego klastra, gdzie *fotoniczne procesory* obsługują najbardziej kosztowne obliczeniowo moduły (np. M1 percepcja CapsNet, M3 abstrakcja Transformer, M4 rozumowanie GNN) z *minimalną latencją*, podczas gdy tradycyjna elektronika koordynuje przepływ danych i logikę. Podsumowując, fotonika **spełnia krytyczne wymagania**: bardzo niską latencję, wysoką przepustowość (zwłaszcza dla obciążeń macierzowych), i potencjalnie lepszą efektywność energii od H100 [2]. Jej *programowalność* staje się coraz lepsza – udowodniono, że nie trzeba specjalnie trenować modeli pod fotoniki, bo układ radzi sobie z natywnymi wagami FP32 [3]. *Skalowalność* również jest obiecująca dzięki łączeniu wielu chipów w jeden pakiet (Lightmatter upakował **6 chipów fotonicznych + kontroler** w jednym module, łącznie **50 mld tranzystorów i 1 mln elementów fotonicznych** [7]). *Dostępność*: technologia jest na etapie prototypów – pierwsze systemy (np. 8-procesorowy serwer Lightmatter) trafiły do laboratoriów [8], lecz masowa dostępność komercyjna może być kwestią najbliższych ~2 lat.

## Architektury neuromorficzne i analogowe

Biologiczny mózg inspiruje architektury, które **przetwarzają informacje poprzez rozproszone, równoległe sumowanie sygnałów (superpozycję prądów)** na neuronach. Dwa podejścia w tej kategorii to: **(a)** analogowe układy pamięciowe (np. memrystory) realizujące sumy i mnożenia macierzowe w sposób ciągły oraz **(b)** asynchroniczne układy *neuromorficzne* symulujące sieci spiking neural networks (SNN). W obu przypadkach unikamy globalnego zegara i sekwencyjnych operacji typowych dla cyfrowych procesorów – zamiast tego wyniki emergentnie powstają z fizycznej dynamiki układu.

**Przykład (a) – analogowe macierze krzyżowe:** Startup **Rain Neuromorphics** zbudował analogowy akcelerator oparty o *64×64* matryce memrystorów, na którym przeprowadzono uczenie sieci rozpoznającej pismo Braille'a [9]. Wyniki są imponujące: czas trenowania zmniejszono o **2 rzędy wielkości** (do dziesiątek **µs**) przy redukcji energii o **5 rzędów** (do setek **nJ**) w porównaniu do GPU [10]. Oznacza to, że pewne operacje, które na GPU trwałyby np. milisekundy, na układzie analogowym dokonują się w dziesiątkach mikrosekund – właśnie dzięki temu, że prądy sumują się jednocześnie w całym crossbarze (superpozycja analogowa), zamiast iteracyjnego mnożenia i sumowania cyfrowego. Rain opracował też algorytmy treningu odporne na szum i niedokładności analogowe (tzw. **MADM – memristor activity-difference method**), omijając problem niekompatybilności standardowego backpropagation z analogowymi rozrzutami parametrów [11]. Co ważne, osiągnięto dokładność taką jak klasyczne algorytmy, co dowodzi wykonalności trenowania głębokich modeli w analogu [12] [13]. **Potencjał skalowania**: Rain projektuje *tile-owalną* architekturę analogową, docelowo zdolną pomieścić 10^11 parametrów (100 mld) i wspierać *ciągłe uczenie się* na krawędzi [14]. Gdyby taka technologia dojrzała, mogłaby zastąpić wiele GPU – wyobraźmy sobie *„analogową kartę AI"* o mocy wielu petaflopsów przy ułamku zużycia energii. To szczególnie atrakcyjne dla **modułów pamięci i uczenia w TITANS** (np. M2 długotrwała pamięć, konsolidacja VAE, M5 rdzeń RL uczący się w pętli meta-kognitywnej) – analogowy układ mógłby na bieżąco aktualizować wagi modeli z minimalnym kosztem energetycznym, umożliwiając **prawdziwie ciągłe uczenie się agenta**.

**Przykład (b) – spiking/neumorficzne układy:** Intel rozwinął rodzinę chipów **Loihi**, które implementują asynchroniczne neurony wyzwalające impulsy (*spike*) po przekroczeniu progu potencjału. Takie układy

przenoszą obliczenia na domenę czasową – informacje są zakodowane w częstości i koincydencji impulsów. *Zaletą* jest znów **niska moc i opóźnienie**: brak cykli zegarowych oznacza, że reakcja systemu następuje natychmiast po dotarciu odpowiednich sygnałów do wyjściowego neuronu. Ponadto, Loihi integruje pamięć z jednostkami obliczeniowymi (każdy „neuron" posiada lokalne przechowanie wag), *minimalizując transfery danych* [15], co dramatycznie obniża zużycie energii. W zadaniach takich jak np. wyszukiwanie najkrótszej ścieżki czy optymalizacja kombinatoryczna, układy neuromorficzne pokazały kilkuset- a nawet tysiąckrotne oszczędności energii względem CPU/GPU, często dorównując im czasem rozwiązania. W kontekście TITANS neuromorficzność mogłaby znaleźć zastosowanie np. w *modułach wykrywania zaskoczenia/anomalii* lub *predykcji następstw* (SNN świetnie radzą sobie z danymi czasowymi i wyciąganiem spatio-temporalnych wzorców [16] [17]). Można sobie wyobrazić **hybrydowy model**: część percepcyjna działa na konwolucyjnej sieci analogowej, potem informacje są kodowane w spiki i przekazywane do sieci SNN, która w sposób ciągły monitoruje sekwencje zdarzeń i wyzwala akcje agentowi (M5) – podobne próby łączenia ANN z SNN już trwają, by łączyć *dokładność* sieci klasycznych z *wydajnością energetyczną* spikingów [18] [19].

**Wyzwania:** Architektury neuromorficzne są **trudniejsze w programowaniu** – wymagają nowych algorytmów trenowania (np. propagacja sygnału przez czas, metody STDP/Equilibrium Prop) i często specjalistycznej wiedzy. Obecnie ich *dostępność* jest ograniczona – Loihi udostępniany jest głównie w środowisku badawczym Intel Neuromorphic Research Community, Rain dopiero planuje komercyjny chip. *Skalowalność* jednak z roku na rok rośnie: **Loihi 2** ma ~1 mln neuronów i możliwość komunikacji wielu chipów przez specjalną sieć, Rain proponuje architekturę 3D z tysiącami matryc memrystorowych. W porównaniu do H100 czy TPU, neuromorficzne podejście **zdecydowanie wygrywa w efektywności energetycznej** (działając nawet *1000× oszczędniej* przy porównywalnej szybkości na pewnych zadaniach) [10], a także może oferować *niższą latencję dla ciągłych strumieni danych* (bo nie czeka na przetworzenie całej „paczek" danych, reaguje event-driven). Jednak **logiczna złożoność** aplikacji, które można na tym uruchomić, bywa ograniczona – nie każdą część pipeline TITANS łatwo zaimplementować jako SNN czy analogowy crossbar. Np. duże modele językowe w obecnej postaci nie działają natywnie w formie spiking (choć trwają prace nad konwersją sieci Transformer do SNN). Dlatego neuromorficzne układy prędzej *uzupełnią* niż całkowicie zastąpią GPU w najbliższym czasie, służąc do przyspieszenia określonych komponentów (np. detekcji zdarzeń, lokalnego uczenia, optymalizacji) w ramach szerszej architektury.

## Układy kwantowe (QPU) – superpozycja stanów kwantowych

Kolejnym podejściem jest wykorzystanie **komputerów kwantowych** – w nich *superpozycja sygnałów* przyjmuje najbardziej dosłowną formę superpozycji stanów kwantowych. QPU (Quantum Processing Unit) operuje na kubitach, które mogą jednocześnie reprezentować kombinacje 0/1, co (dla pewnych algorytmów) pozwala uzyskać **równoległość obliczeń w przestrzeni Hilberta** przewyższającą możliwości klasycznego hardware. W kontekście PIAST-Q jest to szczególnie istotne, bo PIAST-Q to **trapped-ion quantum computer** zintegrowany z infrastrukturą EuroHPC [20]. Ten 20-kubitowy system oferuje *„all-to-all connectivity"* między kubitami i *długie czasy koherencji*, co umożliwia wykonywanie dość złożonych obwodów kwantowych [21]. Zaletą architektur jak PIAST-Q (technologia firmy **AQT**) jest wysoka wierność operacji (fidelity pojedynczych bramek ~99,99% [21]) i możliwość skalowania poprzez dołączanie kolejnych pułapek jonowych (system mieści się w 2 szafach rack [22]). Firma **PsiQuantum** z kolei rozwija **fotoniczny komputer kwantowy** – w pełni oparty o światłowody i fotodetektory – celując w skalę miliona kubitów z korekcją błędów. Już ogłosili chipset *Omega* do fotonicznych kubitów oraz proces produkcyjny w technologii CMOS dla optyki [23] [24].

**Potencjał dla TITANS:** Należy uczciwie zaznaczyć, że *dzisiejsze* komputery kwantowe **nie nadają się do bezpośredniego uruchamiania dużych sieci neuronowych ani LLM** – 20 czy nawet 1000 kubitów to za mało, by zakodować parametry modelu o setkach milionów wag. Ponadto, kwantowe obwody nie

wykonują z natury długich precyzyjnych obliczeń arytmetycznych (są raczej dobre w *algorytmach specjalizowanych* – faktoryzacja, wyszukiwanie nieustrukturyzowane, symulacje fizyczne). Natomiast **możliwe jest hybrydowe użycie QPU jako koprocesora** do pewnych zadań: EuroHPC planuje sprzęgnąć PIAST-Q z superkomputerem klasycznym (najpierw Altair, potem PIAST-AI) właśnie w celu realizacji *hybrydowych algorytmów HPC+QC* [20]. W ramach TITANS/PIAST-Q można by eksperymentować np. z: - **Przyspieszeniem algorytmów optymalizacji**: Rdzeń agentowy (M5) mógłby formułować część problemu decyzyjnego jako QUBO/ising i rozwiązywać go na *kwantowym annealerze* lub przez algorytm wariacyjny (VQE/QAOA). Istnieją prace używające do tego *D-Wave* (choć to kubity typu annealing, nie uniwersalne). - **Losowość i interferencja kwantowa do wyboru modelu**: „Wielomodelowy arbiter LLM" mógłby w teorii korzystać z **kwantowej superpozycji stanów** reprezentujących równocześnie różne modele kandydackie, a następnie poprzez pomiar realizować *losowy wybór z wagami* (to luźna inspiracja – coś jak kwantowy eksplorator przestrzeni modeli). Jednak korzyść nad po prostu losowym wyborem klasycznie jest dyskusyjna. - **Szyfrowanie i prywatność**: QPU mogą służyć do generowania certyfikowanej losowości lub do szybkiego łamania określonych szyfrów – w kontekście bezpieczeństwa TITANS mogłyby pełnić rolę zabezpieczającą kanały komunikacji czy sprawdzającą odporność modelu na pewne ataki (to jednak odległe use-case'y).

Ogólnie, **komputery kwantowe obecnie ustępują klasycznym GPU/TPU w ujęciu ogólnego zastosowania**. Posiadają *dużą latencję* na poziomie całego programu (przygotowanie i odczyt stanu zajmuje milisekundy lub więcej, choć same bramki 1- i 2-kubitowe wykonują się w mikrosekundach). **Przepustowość danych** jest niska – nie przetwarzamy megabajtów, a raczej pojedyncze superpozycje kilkubitowe na raz. Również **efektywność energetyczna** w przeliczeniu na użyteczne operacje nie jest na razie atutem (systemy wymagają chłodzenia kriogenicznego lub laserów i izolacji od drgań). Co istotne, *„komputery kwantowe zmagają się z problemami korekcji błędów, skalowalności i utrzymania koherencji"* [25] – osiągnięcie trwałej przewagi wymaga tysięcy logicznych kubitów bez błędów, co jest celem dopiero na kolejne lata. **Programowalność** QPU to kolejne wyzwanie: tworzenie algorytmów kwantowych jest trudne, a dowiedzenie ich przewagi nad klasycznymi nie zawsze możliwe [25].

**Podsumowanie roli QPU:** W perspektywie projektu PIAST-Q integracja komputera kwantowego jest wartościowa głównie badawczo. Może umożliwić przetestowanie *hybrydowych algorytmów AI* – np. trenowanie małych modeli na zasadzie quantum kernel learning czy testowanie kwantowych wariantów sieci Hopfielda do pamięci asocjacyjnej. Jednak **nie zastąpi klastrów H100/TPU** dla głównego obciążenia, jakim jest trenowanie i inferencja dużych sieci. QPU stanowią raczej *uzupełnienie* klasycznej infrastruktury (co zresztą EuroHPC formalnie realizuje, udostępniając PIAST-Q użytkownikom HPC [20]). Ich *dostępność* jest ograniczona – PIAST-Q ma być gotowy do użytku końcem 2025 [26] dla wybranych użytkowników, komputery PsiQuantum jeszcze nie istnieją w pełnej skali, a np. dostęp do chipów IBM/IONQ jest tylko przez chmurę. Z punktu widzenia TITANS, *logiczna złożoność* systemu AI wykracza poza możliwości obecnych kubitów, więc QPU nie spełni wymagań wydajnościowych (ale warto monitorować postępy, bo pewne nisze obliczeń mogą z czasem osiągnąć *przewagę kwantową*).

## Programowalne układy FPGA i specjalizowane ASIC (Tenstorrent i inne)

Oddzielną kategorię stanowią **układy cyfrowe alternatywne wobec GPU**, które jednak nie operują w analogowej superpozycji sygnałów, lecz optymalizują obliczenia cyfrowe poprzez równoległość i specyficzną architekturę. Wspominamy o nich, bo również mogą obniżyć latencję i koszty:

- **FPGA (Field Programmable Gate Array):** to matryce programowalnych bramek, które pozwalają tworzyć *dedykowane układy logiczne* pod konkretny algorytm. Choć sygnały są tu dwustanowe, FPGA umożliwia np. zaimplementowanie całej sieci neuronowej „w logice sprzętowej" – od

wejścia do wyjścia – bez potrzeb wykonywania programu krok po kroku. Przykładowo Microsoft w projekcie *BrainWave* pokazał inferencję sieci *ResNet-50* na FPGA z opóźnieniem poniżej 1 ms (przy batch size = 1), co było lepsze niż GPU w tak niskim batchu. **Zaletą FPGA jest ultraniski czas propagacji przez pipeline** (każda warstwa jest sprzętowo zrównoleglona) oraz możliwość pracy w stałej arytmetyce całkowitej lub nawet logice binarnej, co redukuje obciążenie. Dla TITANS można by wyobrazić podział: M1 (CapsNet) na jednym FPGA, M2 (VAE pamięć) na drugim, M3 (Transformer) na trzecim, itp., wszystkie połączone szybką siecią. W ten sposób *każdy moduł działa równolegle*, a latencja całego cyklu kognitywnego to tylko suma opóźnień kolejnych modułów (być może < kilka ms), bez kolejkowania zadań na jednym GPU. **Wyzwania**: Niestety, FPGA o wysokiej pojemności (np. Xilinx UltraScale+) są kosztowne i trudne w programowaniu – wymagają napisania opisu sprzętu (HDL) lub użycia narzędzi HLS. Ich *przepustowość* też bywa niższa od GPU przy dużych rozmiarach modeli, bo częstotliwość pracy to ~200-400 MHz, a liczba zasobów logiki ograniczona (dużego transformera nie zmieścimy w jednym układzie bez odwołań do wolniejszej DRAM). *Skalowalność* klastrów FPGA jest ograniczona – istnieją co prawda rozwiązania typu HPC z FPGA, ale brakuje uniwersalnych ekosystemów programistycznych (w przeciwieństwie do CUDA dla GPU). **Programowalność** jest najniższa spośród rozwiązań – każda zmiana modelu może wymagać ponownej syntezy układu (trwającej godziny). Dlatego FPGA są sensowne głównie do *specyficznych zadań w stałej konfiguracji*, gdzie wymagana jest ekstremalnie niska latencja (np. algorytmy high-frequency trading, wnioskowanie małych modeli na krawędzi). W TITANS/PIAST-Q raczej nie są kandydatem do głównego silnika, choć mogą pełnić rolę przyspieszaczy wybranych funkcji (np. FPGA do szybkiego przeszukiwania grafu wiedzy czy realizacji niestandardowych operacji logicznych we wnioskowaniu symboliczno-subsymbolicznym).

- **Tenstorrent (RISC-V AI)**: Tenstorrent to startup (ostatnio mocno dofinansowany przez Samsunga) tworzący *specjalizowane układy ASIC* do AI z rdzeniami **RISC-V**. Ich architektura stawia na **modularność (chiplet)** i otwartość: otwarto opisy interfejsów, a przejęcie firmy **Blue Cheetah Analog** sugeruje nacisk na wydajne analogowe bloki IP w ich układach [27]. Procesory Tenstorrent (np. **Grayskull**) zawierają dziesiątki małych rdzeni Tensix w sieci NoC, każdy z lokalną pamięcią SRAM – to trochę jak GPU podzielone na mniejsze wyspy obliczeń. Wydajność w porównaniu: Grayskull e150 (150 W) osiąga ok. **315 TOPS INT8** (czyli ~78 TFLOPS FP16) [28] [29], co stawia go poniżej flagowych GPU (H100 ~1000 TFLOPS FP8), ale porównywalnie z GPU poprzedniej generacji. Tenstorrent jednak szybko rozwija następców i zapewnia, że ich układy są *łatwe do skalowania w duże klastry*. Otwartość RISC-V daje swobodę dopasowywania formatu liczb, operacji (np. wsparcie dla nietypowych modeli z logiką warunkową czy grafowych, co w GPU by wymagało sporo niestandardowego kodu). Dla TITANS, którego architektura może obejmować zarówno *obliczenia tensorowe* (uczenie, percepcja), jak i *sterowanie logiką agentową* (warunkowe gałęzie, planowanie), posiadanie w jednym układzie uniwersalnych rdzeni CPU *plus* akceleratory może być korzystne. **Latencja**: Tenstorrent wskazuje, że umieszczenie pamięci *on-chip* (np. trzymanie macierzy uwagi Transformera w SRAM) redukuje opóźnienia dostępu i przyspiesza warstwę self-attention kilkukrotnie [30]. Ich układy pracują na dość wysokich zegarach (1 GHz) jak na AI i mogą utrzymywać niskie opóźnienia przy małych partiach danych – to ważne dla interaktywnego agenta. **Energia i koszt**: celem firmy jest **tańsze rozwiązanie niż GPU** poprzez wykorzystanie otwartego IP i skalę produkcji (Samsung jako partner). Dev-kity Grayskull sprzedawano za ~600 USD, co jest ułamkiem ceny GPU – oczywiście z niższą wydajnością, ale stosunek *wydajność/cena* może być korzystny. W zastosowaniach chmurowych Google TPU wykazują ~20% kosztu operacji w porównaniu do GPU [31], a Tenstorrent prawdopodobnie celuje w podobną oszczędność dla własnych klientów. **Programowalność**: Tenstorrent wspiera popularne frameworki (PyTorch, ONNX), a dzięki RISC-V developerzy mogą pisać własny kod niskopoziomowy; to znacznie łatwiejsze niż programowanie FPGA czy neuromorficznego spikera. **Skalowalność**: poprzez łączenie wielu chipletów (architektura *Open*

*Chiplet Architecture* – OCA), możliwe jest zbudowanie modułowych systemów typu *pod* z dziesiątkami układów. W odróżnieniu od np. GPU NVLink, OCA jest otwarte – można by potencjalnie budować europejskie akceleratory z rodzimymi usprawnieniami. **Dostępność**: Pierwsze produkty Tenstorrent (Grayskull, następnie Wormhole) są dostępne dla developerów, ale nie osiągnęły jeszcze masowej adopcji. W najbliższych 1-2 latach mogą jednak pojawić się w centrach danych jako alternatywa, zwłaszcza że ich **wydajność ma rosnąć 2-3x co generację**, podobnie jak TPU Google (TPU v5 ~2× szybsze od v4) [32].

Podsumowując, **specjalizowane układy cyfrowe (ASIC)** jak Tenstorrent oferują *większą elastyczność niż FPGA* i *łatwość integracji*, zachowując wiele zalet GPU (programowalność, wsparcie ekosystemu), a jednocześnie mogą obniżyć **koszty i zużycie energii** poprzez lepszą architekturę i brak narzutu niepotrzebnych funkcji (GPU muszą wspierać bardzo szeroki zakres obliczeń, ASIC mogą być zoptymalizowane pod AI). Dla TITANS może to być **najbardziej realistyczna alternatywa krótkoterminowa** – zbudowanie klastra z np. 8× Tenstorrent o łącznej mocy ~600 W, dającego kilkaset TFLOPS, co odpowiada pojedynczemu H100, ale przy potencjalnie niższej cenie i większej kontroli nad platformą.

## Porównanie technologii a wymagania TITANS/PIAST-Q

Poniżej zestawiono omówione rozwiązania pod kątem kluczowych kryteriów: **latencja**, **przepustowość (wydajność)**, **efektywność energetyczna**, **koszt**, **programowalność**, **skalowalność klastra**, **dostępność** oraz **zdolność obsługi złożonej logiki** wymaganej przez architekturę kognitywną. Dla odniesienia ujęto także standardowe rozwiązania (GPU H100 oraz TPU v5):

| **Technologia** | **Latencja** (opóźnienie) | **Przepustowość / Moc oblicz.** | **Energo-efektywność** | **Koszt (sprzętowy)** | **Programowalność** |
|---|---|---|---|---|---|
| **GPU (Nvidia H100)** | Niska dla dużych batchy (ms); może rosnąć przy wielu modułach kolejno. | Bardzo wysoka (~500–1000 TFLOPS per chip [1]). | ~0.5–1 TOPS/W (wysoka, ale spada pod obciążeniem). | Bardzo wysoki (≥ $30k za akcelerator). | Wysoka (CUDA, szeroki ekosystem). |
| **TPU v5 (Google)** | Niska przy trenowaniu TensorFlow; inferencja batchowa. | Bardzo wysoka (**~460 TFLOPS BF16** per chip) [33]. | Bardzo wysoka (Google podaje 1.3–1.7× efektywniejszy od A100). | W chmurze tani w użyciu (Google ~20% kosztu GPU) [31]; niedostępny poza GCP. | Średnia (ograniczona do TF/JAX, nie full custom). |

| Technologia | Latencja (opóźnienie) | Przepustowość / Moc oblicz. | Energo-efektywność | Koszt (sprzętowy) | Programowalność |
|---|---|---|---|---|---|
| **Fotoniczny (Lightmatter)** | **Bardzo niska** (ns–µs zakres na operacje; jednoprzebiegowe MLP). | Wysoka (nowy chip: **65 TOPS @ 78 W** [4]; do petaflopa w systemie 8×). | **Wysoka** (0.8 TOPS/W w prototypie; potencjał >1 TOPS/W) [4]. | Potencjalnie niższy TCO (mniej energii/ chłodzenia); sprzęt prototypowy kosztowny w R&D. | Średnia+ (wymaga specjalnych kompilatorów, ale uruchamia standardowe modele [3]). |
| **Neuromorficzny (analog/SNN)** | **Bardzo niska** dla event-driven (pojedyncze µs reakcje) [10]; brak narzutu batch. | Średnia (trudno przeliczyć na TFLOPS; demo Rain: mała sieć w µs [10]; Loihi2 ~1M neuronów). | **Bardzo wysoka** (nawet 10^3–10^5 × oszczędność energii na zadaniach) [10]. | Nieznany (Rain R&D; Loihi prototyp; potencjalnie tani w masie – prosta analogowa struktura). | Niska (nowe paradygmaty programowania, brak dojrzałych narzędzi). |
| **QPU (Kwantowy)** | Wysoka (komunikacja i odczyt trwają ms; same bramki ~µs). | Bardzo niska ogólnie (20 kubitów = ~1k operacji jednoczesnych). *Teoretyczna* superpozycja 2^n, ale n małe.* | Niska (duże koszty energii na chłodzenie/ lasery vs parę kubitów operacji). | Ekstremalnie wysoki (PIAST-Q to infrastruktura lab; komercyjnie min. miliony $/system). | Bardzo niska (wymaga specjalistów od mechaniki kwantowej, algorytmy niszowe). |
| **FPGA (np. Xilinx)** | Niska (pipeline na mierze µs, brak kolejki GPU). | Średnia (np. ~10 TFLOPS na duży FPGA; ograniczone zasoby). | Średnia (lepsze wykorzystanie sprzętu przy małych zadaniach; gorsza na duże – niższy zegar). | Wysoki (top FPGA > $10k, plus koszty pracy inżynierskiej). | Niska (HDL/HLS programowanie, długi czas dewelopmentu). |
| **ASIC cyfrowy (Tenstorrent)** | Niska (dedykowane rdzenie, on-chip SRAM minimalizuje opóźnienia pamięci [30]). | Wysoka (current: ~80-100 TFLOPS/ chip FP16; roadmap porównywalny z GPU). | Wysoka (arch. RISC-V + optymalizacje daje niższy pobór na operację niż GPU). | Średni (docelowo niższy niż GPU dzięki masowej produkcji i open IP). | Wysoka (standardowe modele, programowalne jak CPU+GPU combo). |

*Uwagi do tabeli:* Należy pamiętać, że technologie „eksperymentalne” (fotonika, analogi, QPU) są w fazie dynamicznego rozwoju – parametry szybko się poprawiają, ale obecnie często **nie osiągają skali**

wymaganej do zastąpienia w całości klastrów GPU. Dlatego realnym scenariuszem jest **architektura hybrydowa**: wykorzystująca *klasyczne akceleratory (GPU/ASIC)* do większości obliczeń, a *specjalizowane układy* do przyspieszenia newralgicznych fragmentów i obniżenia kosztów operacji.

## Podsumowanie i rekomendacje

Badane alternatywy wskazują, że **„latencja superpozycji"** – czyli wykorzystywanie jednoczesnego nakładania się sygnałów (w świetle, prądzie czy falach kwantowych) – może stać się przełomowym czynnikiem przyspieszającym systemy AI nowej generacji. **Układy fotoniczne** już teraz demonstrują możliwość wykonania obliczeń typowych dla TITANS przy znacznie mniejszym opóźnieniu i mocy niż klasyczne GPU [4]. **Układy analogowe i neuromorficzne** obiecują z kolei zrealizowanie ciągłego uczenia i wnioskowania **in-situ**, co idealnie pasuje do koncepcji samo-udoskonalającego się agenta (moduły M1–M8 mogłyby działać bez przerwy, ucząc się na bieżąco z ułamkiem energii dziś potrzebnej do treningu na GPU [10]). **Komputery kwantowe** są perspektywiczne w niszach, ale na razie głównie symboliczne – ich integracja w PIAST-Q to inwestycja w przyszłość, która może zaprocentować, gdy hybrydowe algorytmy AI+QC dojrzeją. **Specjalizowane ASIC jak Tenstorrent** czy przyszłe europejskie procesory AI mogą natomiast w krótkim terminie *uniezależnić projekt od monopolu Nvidii*, obniżając koszty i dając większą kontrolę nad sprzętem (co wpisuje się w strategię UE wzmacniania suwerenności technologicznej [36]).

**Rekomendacja praktyczna:** rozważyć architekturę integrującą **klaster heterogeniczny** – np. klasyczny superkomputer GPU/CPU (dla ogólnej wydajności i kompatybilności), wsparty **akceleratorami fotonicznymi** dla krytycznych ścieżek neuronowych o dużej przepustowości, oraz **modułem neuromorficznym** obsługującym ciągłe monitorowanie i uczenie (np. nadzorując zmiany w środowisku, wykrywając nowość w strumieniu danych). PIAST-Q jako węzeł kwantowy mógłby być używany on-demand do specyficznych zadań optymalizacyjnych. Takie podejście wpisuje się w globalny trend łączenia wielu technologii – jak ujął to dyrektor EuroHPC: *„PIAST-Q zapewnia dostęp do hybrydowej architektury klasycznej-kwantowej"* [20], a podobnie hybrydy klasyczna-fotoniczna czy klasyczna-neuromorficzna mogą stać się przewagą TITANS.

W dłuższej perspektywie, kontynuując prace badawcze i śledząc rozwój opisanych projektów (Lightmatter, Rain, Loihi, PsiQuantum, Tenstorrent i in.), projekt TITANS może stać się **pionierem wykorzystania nowych architektur sprzętowych** w AI. To nie tylko zwiększy wydajność (niższa latencja, większa przepustowość M1–M8), ale też *obniży koszty operacyjne* (mniej energii = tańsza eksploatacja) oraz *podniesie unikalność projektu* – tworząc agentów AI działających na nietypowych, inspirowanych neurobiologią i fizyką platformach, co idealnie współgra z jego założeniem innowacyjności i przewagi pierwszeństwa.

**Źródła:** Szczegółowe dane i przykłady zaczerpnięto z dokumentacji projektów (Lightmatter [4] [2], Rain Neuromorphics [10], EuroHPC PIAST-Q [21] [20], Tenstorrent [28] oraz analizy własnej infrastruktury TITANS). Powyższa analiza dowodzi, że istnieją realne kierunki rozwoju sprzętowego, które mogą stanowić **alternatywę wobec klastrów GPU H100**, choć wymagają one dalszego dopracowania i integracji. Warto już teraz uwzględnić je w roadmapie projektu – tak, aby TITANS/PIAST-Q był przygotowany na erę post-Moore'owską, w której *superpozycja sygnałów* stanie się kluczem do przełamywania barier wydajności w AI. [25] [10]

---

[1] Comparing NVIDIA's B200 and H100: What's the difference? - Civo
https://www.civo.com/blog/comparing-nvidia-b200-and-h100

2 3 4 5 7 8 25 Photonic AI Acceleration - A New Kind of Computer - Lightmatter®
https://lightmatter.co/blog/a-new-kind-of-computer/

6 Lightmatter Announces Passage L200, the Fastest Co-Packaged ...
https://lightmatter.co/press-release/lightmatter-announces-passage-l200-the-fastest-co-packaged-optics-for-ai/

9 10 11 12 13 14 Rain Demonstrates AI Training on Analog Chip - EE Times
https://www.eetimes.com/rain-demonstrates-ai-training-on-analog-chip/

15 Integrated algorithm and hardware design for hybrid neuromorphic ...
https://www.nature.com/articles/s44335-025-00036-2

16 17 18 19 Integrated algorithm and hardware design for hybrid neuromorphic systems | npj Unconventional Computing
https://www.nature.com/articles/s44335-025-00036-2?error=cookies_not_supported&code=a4317807-616d-4543-a2fa-774fb8ec462c

20 21 26 35 Inauguration of PIAST-Q: A Leap for European Quantum Computing - EuroHPC JU
https://www.eurohpc-ju.europa.eu/inauguration-piast-q-leap-european-quantum-computing-2025-06-23_en

22 AQT to deliver rack-based PIAST-Q quantum computer to PSNC - DCD
https://www.datacenterdynamics.com/en/news/aqt-to-deliver-rack-based-piast-q-quantum-computer-to-psnc/

23 PsiQuantum Announces Omega, a Manufacturable Chipset for ...
https://www.psiquantum.com/news-import/omega

24 PsiQuantum Unveils Omega Chipset for Scalable Photonic Quantum ...
https://quantumcomputingreport.com/psiquantum-unveils-omega-chipset-for-scalable-photonic-quantum-computing/

27 Tenstorrent Acquires Blue Cheetah Analog Design
https://tenstorrent.com/en/vision/tenstorrent-acquires-blue-cheetah-analog-design

28 Tenstorrent Shares Roadmap of Ultra-High-Performance RISC-V ...
https://www.tomshardware.com/news/tenstorrent-shares-roadmap-of-ultra-high-performance-risc-v-cpus-and-ai-accelerators

29 Tenstorrent unveils Grayskull, its RISC-V answer to GPUs
https://news.ycombinator.com/item?id=39658787

30 Attention in SRAM on Tenstorrent Grayskull
https://tenstorrent.com/en/vision/attention-in-sram-on-tenstorrent-grayskull

31 Google's TPU Advantage vs. OpenAI's Nvidia Tax - Nasdaq
https://www.nasdaq.com/articles/cost-ai-compute-googles-tpu-advantage-vs-openais-nvidia-tax

32 Tensor Processing Unit - Wikipedia
https://en.wikipedia.org/wiki/Tensor_Processing_Unit

33 GPU and TPU Comparative Analysis Report | by ByteBridge - Medium
https://bytebridge.medium.com/gpu-and-tpu-comparative-analysis-report-a5268e4f0d2a

34 Lightmatter launches photonic chips to eliminate GPU idle time in AI ...
https://www.networkworld.com/article/3951672/lightmatter-launches-photonic-chips-to-eliminate-gpu-idle-time-in-enterprise-ai-data-centers.html

36 Analiza projektu TITANS.txt
file://file-SwKg53pKPVWQK83168eZ3C