Proposta Técnica e Blueprint Arquitetural: Computação Neuromórfica de Alta Eficiência, Versão 6 (NCHE v6)

Sumário Executivo: Uma Nova Época em Inteligência Artificial com a Arquitetura NCHE v6

Este documento apresenta a proposta técnica e o blueprint arquitetural para a Computação Neuromórfica de Alta Eficiência, Versão 6 (NCHE v6). A NCHE v6 representa uma mudança de paradigma fundamental em relação aos sistemas de inteligência artificial (IA) contemporâneos, afastando-se das arquiteturas convencionais de von Neumann que estão a atingir os seus limites físicos em termos de consumo de energia e estrangulamento de dados. Em vez de ser uma melhoria incremental, a NCHE v6 é uma reimaginação holística da computação, concebida de raiz para emular os princípios fundamentais de eficiência, adaptabilidade e robustez que governam a computação biológica. O seu objetivo é servir como a plataforma fundacional para a próxima geração de IA — sistemas que são verdadeiramente autónomos, capazes de aprendizagem contínua ao longo da vida e que operam com uma eficiência energética sem precedentes.

As principais inovações da arquitetura NCHE v6 são multifacetadas e profundamente integradas:

1. **Núcleo Computacional Bio-plausível:** No coração da NCHE v6 está o neurónio de Izhikevich, um modelo que equilibra de forma única a riqueza dinâmica do comportamento neuronal biológico com a eficiência computacional.[1] Esta escolha permite que a rede exiba um repertório complexo de padrões de disparo temporal, essencial para a computação sofisticada.
2. **Substrato Físico de Vanguarda:** A arquitetura é fisicamente realizada através de uma integração monolítica 3D (M3D) que funde três tecnologias díspares numa única matriz: lógica de controlo CMOS, sinapses de memristores de óxido de háfnio (HfO_2) ferroelétricos e uma Rede Ótica em Chip (ONoC) de fotónica de silício.[3] Esta integração vertical elimina o estrangulamento da memória,

colocando a computação, a memória e a comunicação em proximidade nanométrica, reduzindo drasticamente a latência e o consumo de energia.

3. **Motor de Aprendizagem Unificado e Multi-fator:** A NCHE v6 abandona as regras de aprendizagem simplistas em favor de um motor de plasticidade unificado que espelha a complexidade da aprendizagem biológica. Este motor combina: (a) Plasticidade Dependente do Tempo do Spike (STDP) probabilística e consciente do hardware, que abraça a estocasticidade dos dispositivos memristivos como uma característica computacional [6]; (b) STDP modulada por recompensa (R-STDP), que permite a aprendizagem por reforço orientada por objetivos através de sinais de "terceiro fator" neuromoduladores [8]; e (c) Plasticidade Homeostática e Estrutural (HSP), que regula a atividade da rede e permite a sua reorganização física (crescimento e poda de sinapses) para estabilidade e aprendizagem ao longo da vida.[9]
4. **Inteligência ao Nível do Sistema:** Para além do hardware, a NCHE v6 incorpora inteligência ao nível do sistema. Isto inclui um co-design de algoritmo-hardware para mapeamento e otimização eficientes [11], o uso de algoritmos genéticos de inspiração quântica (QIGA) para otimização da topologia da rede e dos hiperparâmetros de aprendizagem [13], mecanismos de auto-reparação inspirados nos astrócitos para tolerância a falhas [15] e uma estrutura de IA Explicável (XAI) concebida para a dinâmica temporal das redes neuronais de spiking.[16]

As projeções de desempenho, baseadas em investigações de vanguarda sobre os componentes individuais, indicam que a NCHE v6 alcançará ganhos de eficiência energética superiores a 100 vezes em comparação com as soluções baseadas em GPU para cargas de trabalho de IA equivalentes [17], ao mesmo tempo que permite capacidades de aprendizagem contínua em tempo real em ambientes dinâmicos.[18] Ao abordar os desafios fundamentais de energia, adaptação e complexidade, a NCHE v6 não se destina a ser apenas um acelerador mais rápido, mas sim o catalisador para uma nova era de IA — uma IA que pode aprender, adaptar-se e operar de forma eficiente e segura no mundo real.

Parte I: Princípios Fundamentais da NCHE v6

1.1. Filosofia Central: Emular Princípios Biológicos de Computação

A filosofia de design da arquitetura NCHE v6 representa uma transição fundamental de uma inspiração metafórica do cérebro para a emulação de princípios de engenharia concretos que governam a computação biológica. Gerações anteriores de hardware de IA, embora bem-sucedidas nos seus domínios, permaneceram em grande parte confinadas pela arquitetura de von Neumann, onde a separação física entre as unidades de processamento e de memória cria um estrangulamento fundamental no movimento de dados — um problema que consome a maior parte da energia nos sistemas de IA modernos.[15] A NCHE v6 é concebida desde o início para demolir esta "parede da memória" e abraçar os princípios que tornam o cérebro um computador tão eficiente.

Os princípios centrais que orientam o design da NCHE v6 são:

1. **Processamento Orientado a Eventos:** A computação só ocorre quando um evento significativo — um "spike" — acontece. Ao contrário dos sistemas síncronos baseados em relógio que processam continuamente dados (muitas vezes redundantes), a natureza assíncrona da NCHE v6 garante que a energia é consumida apenas quando nova informação está a ser transmitida.[20] Isto leva a uma representação esparsa da atividade, onde apenas um pequeno subconjunto de neurónios está ativo a qualquer momento, espelhando a esparsidade observada no cérebro e contribuindo para uma eficiência energética drástica.[20]
2. **Co-localização da Memória e Computação (Computação na Memória):** No cérebro, a sinapse é simultaneamente a unidade de memória (armazenando a força da conexão) e a unidade de computação (ponderando os sinais de entrada). A NCHE v6 emula este princípio ao utilizar matrizes de memristores como sinapses, onde o peso sináptico (resistência) é armazenado e a multiplicação do vetor da matriz (a operação fundamental da rede neuronal) é realizada *in situ* através das leis da física (Lei de Ohm e Leis de Kirchhoff).[22] Isto elimina a necessidade de transferir pesos da memória para uma unidade de processamento separada, o passo mais consumidor de energia na computação de IA convencional.
3. **Paralelismo Massivo:** A arquitetura é inerentemente paralela, com milhares de núcleos neuromórficos a operar simultaneamente e de forma independente. A comunicação entre estes núcleos é gerida por uma rede em chip de alta largura de banda, permitindo a escala para redes com milhões de neurónios e milhares de milhões de sinapses, semelhante à escala do cérebro.[24]

Os objetivos de design que emanam desta filosofia são claros e ambiciosos. O

principal objetivo é alcançar uma melhoria de ordens de magnitude na eficiência energética, com uma meta de mais de 100 vezes os ganhos em relação às Unidades de Processamento Gráfico (GPUs) para cargas de trabalho de IA comparáveis, conforme sugerido por estudos sobre hardware neuromórfico existente.[17] O segundo objetivo principal é permitir a aprendizagem contínua ao longo da vida, onde o sistema pode aprender novas tarefas sequencialmente a partir de um fluxo contínuo de dados sem esquecer catastroficamente o conhecimento anterior — uma capacidade crucial para sistemas autónomos que operam em ambientes dinâmicos.[18] Finalmente, a arquitetura é otimizada para o processamento em tempo real de dados de sensores esparsos e orientados a eventos, como os de sensores de visão dinâmica (DVS), tornando-a uma plataforma ideal para aplicações em robótica, veículos autónomos e dispositivos de ponta com restrições de energia.[26] Ao construir o sistema sobre estes princípios biológicos, a NCHE v6 visa ir além da simples aceleração dos algoritmos de IA existentes para permitir uma nova classe de IA adaptativa, robusta e eficiente.

1.2. Visão Geral da Arquitetura: Um Sistema Hierárquico e Multi-escala

A arquitetura NCHE v6 é um sistema profundamente integrado, concebido em múltiplas camadas de abstração, desde o substrato físico até ao plano de controlo de alto nível. Cada camada é co-desenhada para funcionar em sinergia com as outras, formando um todo coeso que encarna os princípios fundamentais da computação neuromórfica.

A pilha arquitetónica da NCHE v6 pode ser visualizada da seguinte forma:

1. **Plano Físico (Substrato):** A base da NCHE v6 é um chip de integração monolítica 3D (M3D-IC). Esta tecnologia de fabrico de ponta permite a empilhamento vertical de camadas de materiais e dispositivos díspares num único substrato de silício. A pilha consiste em: (a) uma camada base de lógica **CMOS** para os circuitos neuronais digitais e controlo do sistema; (b) uma camada intermédia de **memristores de HfO_2** em matrizes de barras cruzadas densas, que funcionam como sinapses analógicas; e (c) uma camada superior de **Fotónica de Silício (SiPh)** que implementa uma Rede Ótica em Chip (ONoC) para comunicação de alta largura de banda e baixa latência.[3]
2. **Plano Computacional (Núcleo):** O núcleo computacional é composto por populações de **neurónios de Izhikevich** programáveis, capazes de uma rica

dinâmica temporal.[1] Cada núcleo neuromórfico contém estes neurónios e as suas sinapses memristivas associadas. Crucialmente, cada núcleo também alberga uma instância do
motor de plasticidade unificado, permitindo que a aprendizagem (p-STDP, R-STDP, HSP) ocorra localmente, com base na atividade local e em sinais de modulação globais.[8]

3. **Plano de Comunicação (Rede):** A comunicação entre os núcleos neuromórficos é gerida pela **Rede Ótica em Chip (ONoC) de Mundo Pequeno Hierárquico**. Spikes são codificados como pulsos de luz e transmitidos através de guias de onda de silício. A Multiplexagem por Divisão de Comprimento de Onda (WDM) permite que um único guia de onda transporte centenas de canais de spikes em paralelo, enquanto micro-ressoadores em anel atuam como comutadores e filtros para encaminhar os spikes para os seus destinos.[25] A topologia geral da rede segue um padrão de mundo pequeno, com clusters locais densamente conectados e atalhos esparsos de longo alcance, espelhando a conectividade eficiente do cérebro.[31]
4. **Plano de Controlo e Programação (Sistema):** A camada mais alta da pilha gere a operação e programação do sistema. Inclui uma **estrutura de co-design de algoritmo-hardware** que mapeia as Redes Neuronais de Spiking (SNNs) para o hardware de forma otimizada.[12] Um
módulo de Otimização de Inspiração Quântica (QIO) utiliza algoritmos clássicos inspirados na computação quântica para otimizar offline a topologia da rede e os hiperparâmetros.[13] Finalmente, uma
interface de IA Explicável (XAI) fornece ferramentas para interpretar as decisões da rede, analisando as cadeias causais de spikes que levaram a um resultado específico.[16]

O fluxo de dados através deste sistema é inerentemente assíncrono e orientado a eventos. Os dados de entrada, idealmente de sensores neuromórficos como câmaras DVS que produzem fluxos de eventos esparsos [35], são codificados em padrões de spikes temporais. Estes spikes propagam-se através da ONoC para os núcleos neuromórficos relevantes. Dentro de cada núcleo, os spikes de entrada são ponderados pelas sinapses memristivas e integrados pelos neurónios de Izhikevich. Quando o potencial de um neurónio atinge o seu limiar, ele dispara um spike de saída, que é então transmitido a outros núcleos através da ONoC. Simultaneamente, as regras de plasticidade locais e globais modificam continuamente os pesos sinápticos e a própria estrutura da rede, permitindo que o sistema aprenda e se adapte em tempo real, sem uma separação entre as fases de "treino" e "inferência". Este fluxo contínuo de computação, comunicação e aprendizagem distribuída é o que confere à

NCHE v6 a sua potência e eficiência.

Parte II: O Núcleo Computacional: Neurónios, Sinapses e Plasticidade

O coração da arquitetura NCHE v6 reside no seu núcleo computacional, onde os princípios da neurociência são traduzidos em modelos matemáticos e implementações físicas. Esta secção detalha os três pilares deste núcleo: a unidade neuronal, a unidade sináptica e o motor de aprendizagem que as une. A seleção e o co-design destes componentes são cruciais para alcançar o equilíbrio desejado entre a plausibilidade biológica, a eficiência computacional e a capacidade de adaptação.

2.1. A Unidade Neuronal: O Modelo de Neurónio de Izhikevich

A escolha do modelo neuronal é uma decisão de design fundamental que influencia profundamente as capacidades computacionais e a eficiência de uma arquitetura neuromórfica. Enquanto modelos simples como o Leaky Integrate-and-Fire (LIF) oferecem alta eficiência computacional, mas carecem de riqueza dinâmica, modelos complexos como o de Hodgkin-Huxley (HH) proporcionam um realismo biológico excecional à custa de um custo computacional proibitivo para simulações em larga escala.[2] A arquitetura NCHE v6 adota o modelo de neurónio de Izhikevich, uma solução que ocupa uma posição ideal neste espectro, oferecendo um compromisso excecional entre eficiência e plausibilidade biológica.[1]

O modelo de Izhikevich é notável pela sua capacidade de reproduzir uma vasta gama de comportamentos de disparo neuronal observados em diferentes regiões do córtex, incluindo padrões como picos regulares (regular spiking - RS), picos intrinsecamente em rajada (intrinsically bursting - IB) e picos em tagarelice (chattering - CH), entre outros.[1] Esta diversidade dinâmica é essencial para a computação temporal complexa e para a formação de oscilações e grupos de disparo policrónicos realistas, que se acredita serem fundamentais para a codificação de informação no cérebro.[2] Apesar desta riqueza, o modelo é computacionalmente tão eficiente quanto um modelo de integração e disparo, tornando viável a simulação de redes em larga escala em tempo

real.[1]

A dinâmica do neurónio de Izhikevich é descrita por um sistema de duas equações diferenciais ordinárias acopladas:

dtdv=0.04v2+5v+140−u+Idtdu=a(bv−u)
Seguido por uma condição de reposição após o disparo:

se v≥30 mV, enta˜o {v←cu←u+d
Nestas equações:

- v representa o potencial da membrana do neurónio.
- u é uma variável de recuperação da membrana, que modela a ativação e desativação das correntes iónicas e proporciona feedback negativo para v.
- I é a corrente sináptica total de entrada para o neurónio.
- a,b,c,d são quatro parâmetros adimensionais que são ajustados para replicar os diferentes padrões de disparo dos neurónios corticais. O parâmetro a descreve a escala de tempo da variável de recuperação u. O parâmetro b descreve a sensibilidade de u às flutuações do potencial de membrana v. O parâmetro c descreve o valor de reposição de v após um pico, e d descreve o valor de reposição de u após um pico.[1]

Na NCHE v6, estes modelos neuronais serão implementados digitalmente em cada núcleo neuromórfico, utilizando tecnologia de processo avançada (semelhante ao Intel 4 usado no Loihi 2 [38]) para garantir a eficiência. Uma característica chave é que os parâmetros

a,b,c,d serão programáveis para cada neurónio ou população de neurónios. Isto permite a criação de redes heterogéneas, onde diferentes tipos de neurónios podem coexistir e desempenhar papéis computacionais distintos, aumentando significativamente o poder expressivo da arquitetura.

A tabela seguinte resume a justificação para a escolha do modelo de Izhikevich em comparação com outros modelos proeminentes.

Tabela 2.1: Análise Comparativa de Modelos de Neurónios de Spiking

Característica	Modelo Hodgkin-Huxley (HH)	Modelo Leaky Integrate-and-Fire (LIF)	Modelo de Izhikevich (NCHE v6)

Plausibilidade Biológica	Muito Alta. Modela a dinâmica detalhada dos canais iónicos.[2]	Baixa. Modelo linear simplificado que não captura a dinâmica neuronal complexa.[2]	Alta. Reproduz mais de 20 padrões de disparo neuronal conhecidos com apenas 2 equações.[1]
Custo Computacional	Muito Alto. Requer a resolução de múltiplas equações diferenciais não lineares.[2]	Muito Baixo. Requer a resolução de uma única equação diferencial linear.[1]	Baixo. Computacionalmente tão eficiente quanto o modelo LIF, permitindo simulações em larga escala.[1]
Riqueza Dinâmica	Muito Alta. Capaz de modelar com precisão os mecanismos biofísicos.	Muito Baixa. Tipicamente exibe apenas um padrão de disparo (integração).	Muito Alta. Capaz de gerar uma vasta gama de comportamentos, incluindo rajadas, tagarelice e adaptação.[37]
Justificação para NCHE v6	Rejeitado devido ao custo computacional proibitivo para sistemas em larga escala.	Rejeitado devido à falta de riqueza dinâmica necessária para a computação temporal complexa.	**Selecionado** por oferecer o melhor compromisso entre plausibilidade biológica e eficiência computacional.

2.2. A Unidade Sináptica: Memristores Ferroelétricos de HfO_2

A sinapse, o ponto de conexão entre neurónios, é onde a aprendizagem e a memória ocorrem. A sua implementação física é, portanto, um dos aspetos mais críticos de qualquer arquitetura neuromórfica. A NCHE v6 utiliza memristores como sinapses analógicas não voláteis, uma escolha motivada pela sua capacidade de realizar a computação na memória e pela sua semelhança funcional com as sinapses biológicas.[23] Especificamente, a NCHE v6 adota memristores ferroelétricos baseados em Óxido de Háfnio (HfO_2) dopado, uma tecnologia que supera muitas das limitações das memórias resistivas de acesso aleatório (RRAM) convencionais baseadas em filamentos.[40]

O desafio com muitos memristores de óxido metálico é a sua dependência da formação e rutura estocástica de filamentos condutores, um processo que leva a uma variabilidade significativa de dispositivo para dispositivo e de ciclo para ciclo, baixa fiabilidade e dificuldade em alcançar estados de resistência multi-nível estáveis.[39] Os memristores ferroelétricos de HfO_2 evitam este problema. O seu mecanismo de comutação não se baseia em filamentos, mas sim na polarização estável de domínios ferroelétricos dentro do material de HfO_2. A aplicação de um campo elétrico pode inverter esta polarização, alterando a altura da barreira de Schottky na interface entre o material ferroelétrico e o elétrodo, o que, por sua vez, modula a resistência do dispositivo.[40]

As principais vantagens que justificam a escolha do HfO_2 para a NCHE v6 são:

- **Estabilidade e Reprodutibilidade Superiores:** Como a comutação é um efeito de campo que governa domínios estáveis, em vez da migração aleatória de iões, estes dispositivos exibem uma variabilidade muito menor, maior resistência e melhor retenção de dados.[41]
- **Capacidade Multi-Nível Analógica:** A resistência do dispositivo pode ser ajustada de forma gradual e contínua, controlando a fração de domínios ferroelétricos que são comutados. Isto permite a implementação de pesos sinápticos analógicos de alta resolução (a NCHE v6 visa >6 bits), o que é crucial para a precisão dos algoritmos de aprendizagem.[40]
- **Alto Desempenho e Eficiência:** Os dispositivos de HfO_2 demonstraram rácios on/off elevados (>10^3), tempos de retenção longos (>10^4 s) e alta resistência (>10^4 ciclos), cumprindo os requisitos para aplicações de computação robustas.[39]
- **Compatibilidade com CMOS:** O HfO_2 já é um material "high-k" padrão na indústria de semicondutores, o que facilita a sua integração em processos de fabrico CMOS existentes, especialmente em configurações de back-end-of-line (BEOL).[4]

Na arquitetura NCHE v6, estas sinapses memristivas serão organizadas em matrizes de barras cruzadas densas com uma configuração 1T1R (um transístor, um resistor). O transístor em cada célula atua como um seletor, permitindo o endereçamento preciso de sinapses individuais para operações de leitura e escrita e mitigando as correntes de fuga ("sneak path") que podem corromper os cálculos em matrizes de barras cruzadas passivas.[19] Estas matrizes formam a base da memória computacional da NCHE v6, onde os pesos são armazenados e as operações de multiplicação de vetor por matriz são executadas de forma eficiente e paralela.

A tabela seguinte compara as propriedades de diferentes tecnologias de memristores

candidatas, solidificando a escolha do HfO_2 ferroelétrico.

Tabela 2.2: Propriedades de Materiais Memristivos Candidatos para Implementação Sináptica

Propriedade	RRAM de Óxido Filamentar (ex: TaOx)	Memória de Mudança de Fase (PCM)	Memristor Ferroelétrico de HfO_2 (NCHE v6)
Estabilidade (Variabilidade)	Baixa a Média. A natureza estocástica da formação/rutura do filamento leva a uma alta variabilidade D2D e C2C.[42]	Média. Sofre de deriva de resistência ao longo do tempo devido ao relaxamento estrutural da fase amorfa.	Alta. A comutação baseada em polarização de domínio estável evita a estocasticidade do filamento, resultando em baixa variabilidade.[41]
Resistência	Média a Alta (10^4 - 10^6 ciclos).[39]	Alta (>10^8 ciclos).	Alta (>10^4 ciclos, com potencial para mais).[39]
Retenção	Boa (>10 anos), mas pode ser afetada pela instabilidade do filamento.[45]	Boa (>10 anos), mas a deriva de resistência é uma preocupação.	Excelente (>10^4 s, com potencial para >10 anos).[41]
Capacidade Multi-Nível	Limitada. Difícil de controlar com precisão o tamanho do filamento para obter estados analógicos estáveis.	Possível, mas a deriva de resistência complica a distinção entre níveis próximos.	Excelente. A polarização parcial permite estados multi-nível estáveis e de alta resolução, ideal para pesos sinápticos analógicos.[40]
Energia/Escrita	Baixa a Média.	Alta. Requer pulsos de alta corrente para derreter o material.	Baixa. A comutação por campo elétrico é inerentemente eficiente em termos de energia.
Justificação para NCHE v6	Rejeitado devido à alta variabilidade e	Rejeitado devido ao alto consumo de	**Selecionado** pela sua estabilidade

	dificuldade em alcançar plasticidade analógica fiável.	energia na escrita e problemas de deriva.	superior, capacidade multi-nível analógica e compatibilidade com CMOS, abordando os principais desafios dos memristores.

2.3. O Motor de Aprendizagem: Uma Estrutura de Plasticidade Unificada e Multi-fator

A capacidade de uma rede neuromórfica aprender e adaptar-se é o que a distingue de um circuito estático. A NCHE v6 implementa um motor de aprendizagem sofisticado que vai muito além das regras de Hebbian simples. É uma estrutura unificada onde múltiplos mecanismos de plasticidade, inspirados em diferentes processos biológicos, operam em conjunto para moldar a rede. Esta abordagem sinérgica é fundamental para alcançar a estabilidade a longo prazo, a aprendizagem orientada por objetivos e a adaptação ao longo da vida. Os três componentes principais deste motor são: Plasticidade Dependente do Tempo do Spike (STDP) probabilística e consciente do hardware, aprendizagem por reforço neuromodulada e plasticidade homeostática e estrutural.

A interação destes três mecanismos cria um sistema de aprendizagem robusto e multifacetado. A p-STDP não supervisionada permite que a rede aprenda autonomamente as características estatísticas do seu ambiente, formando detetores de padrões eficientes.[46] No entanto, a aprendizagem puramente Hebbiana pode ser instável e não é inerentemente orientada para um objetivo.[47] É aqui que a R-STDP entra, fornecendo um sinal de recompensa global que orienta a plasticidade local para reforçar os caminhos e as representações que levam a resultados desejáveis, transformando a aprendizagem não supervisionada em aprendizagem por reforço.[8] Contudo, esta combinação de plasticidade pode levar a uma atividade descontrolada, com alguns neurónios a ficarem hiperativos e outros silenciosos. A HSP atua como o sistema de controlo regulador final. A homeostase da taxa de disparo estabiliza a atividade de neurónios individuais através da escala sináptica [49], enquanto a plasticidade estrutural fornece um mecanismo para a adaptação a longo prazo, podando conexões inúteis e criando novas para explorar o espaço de soluções, permitindo que a rede escape de mínimos locais e se adapte a novas tarefas e ambientes ao longo da sua vida.[10] Esta tríade de aprendizagem — extração de

características (p-STDP), otimização de objetivos (R-STDP) e regulação da estabilidade (HSP) — é o que dota a NCHE v6 da sua capacidade de aprendizagem contínua e ao longo da vida.[18]

2.3.1. STDP Probabilística e Consciente do Hardware (p-STDP)

A abordagem da NCHE v6 à plasticidade sináptica fundamental afasta-se dos modelos idealizados e abraça as realidades físicas do seu substrato memristivo. Em vez de combater a variabilidade e a não idealidade dos memristores de HfO_2, a arquitetura aproveita-as como uma característica computacional através de uma regra de STDP probabilística (p-STDP) e dependente do estado.

Nos modelos de STDP tradicionais, a mudança no peso sináptico (Δw) é uma função determinística da diferença de tempo entre os spikes pré e pós-sinápticos (Δt). No entanto, em dispositivos físicos como os memristores, a mudança na condutância não depende apenas do pulso de programação, mas também do estado atual da condutância do dispositivo.[6] Além disso, o processo de comutação em nanoescala é inerentemente estocástico.[45]

A p-STDP na NCHE v6 modela explicitamente estes dois fenómenos:

1. **Dependência do Estado:** A magnitude da mudança de peso potencial, Δw, é uma função não linear tanto de Δt como do peso atual, w. À medida que w se aproxima dos seus limites superior ou inferior (correspondendo aos estados de baixa e alta resistência do memristor), a magnitude da mudança de peso diminui, modelando o efeito de saturação natural do dispositivo.[53]
2. **Comutação Probabilística:** A atualização do peso não é garantida. Em vez disso, é aplicada com uma probabilidade, Pupdate, que é uma função da tensão efetiva através do memristor gerada pelo mecanismo de STDP. Isto é modelado como uma função sigmoidal, onde pulsos de tensão mais fortes (resultantes de correlações de spikes mais próximas no tempo) têm uma maior probabilidade de superar a barreira de energia para a comutação.[7]

Esta abordagem transforma um potencial "bug" do hardware (variabilidade) numa "característica" algorítmica. A estocasticidade inerente atua como uma forma de regularização, semelhante ao dropout nas redes neuronais artificiais (ANNs), evitando o sobreajuste e melhorando a generalização. Ao alinhar o algoritmo de aprendizagem com a física do hardware subjacente, a NCHE v6 alcança uma forma de co-design de

algoritmo-hardware que promove a robustez computacional.[12]

2.3.2. Aprendizagem por Reforço Neuromodulada (R-STDP)

Para permitir que a rede aprenda a atingir objetivos específicos, a NCHE v6 aumenta a sua regra de p-STDP local com um sinal de "terceiro fator" global, inspirado na ação de neuromoduladores como a dopamina no cérebro.[55] Este mecanismo, conhecido como STDP modulada por recompensa (R-STDP), transforma a aprendizagem Hebbiana não supervisionada num poderoso paradigma de aprendizagem por reforço.[8]

O mecanismo funciona da seguinte forma:

1. **Traços de Elegibilidade:** Quando ocorrem correlações de spikes pré e pós-sinápticas, em vez de induzirem uma mudança de peso imediata, criam um "traço de elegibilidade" de curta duração na sinapse, e(t).[56] Este traço é uma etiqueta temporária que marca a sinapse como candidata a plasticidade.
2. **Sinal de Recompensa Global:** Um sinal de recompensa global, R(t), é transmitido para toda a rede (ou para regiões específicas). Este sinal é gerado externamente ou por um módulo de "crítica" dentro da própria rede, indicando o sucesso ou o fracasso na realização de uma tarefa.
3. **Atualização de Peso Fechada:** A mudança de peso permanente só ocorre quando o sinal de recompensa R(t) interage com o traço de elegibilidade e(t). A regra de atualização torna-se, conceptualmente, $\Delta w \propto R(t) \cdot e(t)$.[57] Um sinal de recompensa positivo ($R>0$) reforçará as mudanças de peso sugeridas pelo traço (aprendizagem Hebbiana), enquanto um sinal de punição ($R<0$) pode suprimir ou mesmo inverter a plasticidade (aprendizagem anti-Hebbiana), encorajando a rede a explorar outras estratégias.

Este mecanismo resolve o problema da atribuição de crédito temporal, ligando as ações locais (correlações de spikes) a resultados globais e muitas vezes atrasados (recompensa da tarefa).[8] Permite que a NCHE v6 aprenda a selecionar ações que maximizam a recompensa ao longo do tempo, uma capacidade fundamental para a tomada de decisões em agentes autónomos.

2.3.3. Plasticidade Homeostática e Estrutural (HSP)

Para garantir a estabilidade a longo prazo e a capacidade de aprendizagem contínua, a NCHE v6 incorpora dois mecanismos reguladores adicionais: a plasticidade homeostática e a plasticidade estrutural.[9]

1. **Plasticidade Homeostática:** As redes com plasticidade Hebbiana são propensas à instabilidade, onde o feedback positivo pode levar a uma atividade descontrolada ou ao silenciamento de neurónios. A plasticidade homeostática combate isto regulando a excitabilidade de cada neurónio para manter uma taxa de disparo alvo.[49] Se a taxa de disparo média de um neurónio, $\langle a \rangle$, se desvia do seu ponto de ajuste, a_{target}, os mecanismos de escala sináptica ajustam a força de todas as suas sinapses de entrada para o trazer de volta ao intervalo de operação desejado. Se $\langle a \rangle > a_{target}$, os pesos de entrada são reduzidos (escala para baixo); se $\langle a \rangle < a_{target}$, os pesos são aumentados (escala para cima).[49]
2. **Plasticidade Estrutural:** Indo além da simples modificação de pesos, a NCHE v6 pode reorganizar fisicamente a sua conectividade. Este processo de plasticidade estrutural é impulsionado pela atividade e pela pressão homeostática.[50] O modelo baseia-se na ideia de que os neurónios têm um número de "elementos sinápticos" (análogos aos botões axonais e espinhas dendríticas) que podem ser criados ou eliminados com base na sua atividade.[10]
    - **Crescimento e Poda:** Se um neurónio está cronicamente subativo, criará novos elementos sinápticos para procurar mais entradas. Se estiver hiperativo, podará os elementos existentes para reduzir a sua carga de entrada.[62]
    - **Formação de Sinapses:** Elementos sinápticos "vagos" (não conectados) podem então formar novas sinapses. Este processo de formação de conexões é estocástico e dependente da distância, favorecendo conexões locais, mas permitindo a formação de conexões de longo alcance com uma probabilidade não nula. Curiosamente, os modelos mostram que, à medida que a rede se aproxima do seu equilíbrio homeostático, começa a formar mais conexões de longo alcance, aumentando a eficiência da rede (tornando-a mais "mundo pequeno").[10]

Esta capacidade de crescer, podar e religar conexões dota a NCHE v6 de uma notável adaptabilidade, permitindo-lhe alocar recursos de forma eficiente e remodelar-se continuamente em resposta a novas informações e tarefas, uma pedra angular da

aprendizagem ao longo da vida.

A tabela seguinte fornece as formulações matemáticas centrais para cada componente do motor de aprendizagem da NCHE v6.

Tabela 2.3: Formulação Matemática das Regras de Plasticidade Multi-fator da NCHE v6

Mecanismo de Plasticidade	Formulação Matemática Central	Descrição das Variáveis
p-STDP	$\Delta w_{ij} = f(w_{ij}, \Delta t)$ com probabilidade P_{update} onde $f(w_{ij}, \Delta t) = \{ A_+ e^{-\Delta t/\tau_+} (w_{max} - w_{ij}) - A_- e^{\Delta t/\tau_-} (w_{ij} - w_{min}) \text{se } \Delta t > 0 \text{se } \Delta t < 0$ $P_{update} = \sigma(V_{pulse} - V_{threshold})$	w_{ij}: peso sináptico atual. $\Delta t = t_{post} - t_{pre}$: diferença de tempo do spike. A_+, A_-: amplitudes de aprendizagem. τ_+, τ_-: constantes de tempo da janela de STDP. w_{max}, w_{min}: limites de peso do memristor. P_{update}: probabilidade de atualização. σ: função de comutação estocástica.
R-STDP	$\frac{dw_{ij}}{dt} = R(t) \cdot e_{ij}(t)$ $\frac{de_{ij}}{dt} = -\frac{e_{ij}}{\tau_e} + \text{STDP\_events}(\Delta t)$	$R(t)$: sinal de recompensa global (escalar). $e_{ij}(t)$: traço de elegibilidade sináptico. τ_e: constante de tempo de decaimento da elegibilidade. STDP_events: contribuições da p-STDP para o traço.
HSP	Escala Sináptica: $\frac{dw_{ij}}{dt} = \eta_h(\langle a_j \rangle - a_{target}) w_{ij}$ Plasticidade Estrutural: $P(\text{criação}) \propto (a_{target} - \langle a_j \rangle)$ $P(\text{poda}) \propto (\langle a_j \rangle - a_{target})$	η_h: taxa de aprendizagem homeostática. $\langle a_j \rangle$: taxa de disparo média do neurónio pós-sináptico j. a_{target}: taxa de disparo alvo. $P(\cdot)$: probabilidade de criação/poda de um elemento sináptico.

Parte III: O Substrato Físico: Integração Monolítica e

Comunicação

A realização física da arquitetura NCHE v6 exige uma abordagem radicalmente diferente da fabricação de chips 2D convencional. Para alcançar a densidade, velocidade e eficiência energética sem precedentes exigidas pela sua filosofia de design, a NCHE v6 depende de duas tecnologias de hardware de vanguarda: a integração monolítica 3D e uma rede ótica em chip. Esta combinação permite a fusão de tecnologias de dispositivos díspares — lógica CMOS, memória memristiva e fotónica de silício — num único sistema coeso, resolvendo fundamentalmente o estrangulamento de comunicação que assola a computação de alto desempenho.

3.1. Integração Monolítica 3D: Fundindo Memristores, Fotónica e Lógica de Controlo

A integração monolítica 3D (M3D) é uma técnica de fabrico avançada onde múltiplas camadas de dispositivos são construídas sequencialmente numa única bolacha de silício.[3] Ao contrário das abordagens de empilhamento 3D paralelas que ligam chips fabricados separadamente através de vias de silício (TSVs), a M3D utiliza vias de interligação entre camadas (ILVs) de escala nanométrica definidas por litografia. Isto permite uma densidade de interconexão vertical que é ordens de magnitude superior, permitindo uma comunicação de latência ultra-baixa e alta largura de banda entre as diferentes camadas funcionais.[3]

A arquitetura da pilha M3D da NCHE v6 é concebida da seguinte forma:

1. **Camada Inferior (Lógica CMOS):** A camada base é fabricada utilizando um processo CMOS padrão e maduro. Esta camada contém os circuitos digitais que implementam os neurónios de Izhikevich, a lógica de controlo para os mecanismos de plasticidade, os controladores de memória para as matrizes de barras cruzadas e a interface para o plano de programação do sistema. A utilização de um processo CMOS estabelecido garante um alto rendimento e fiabilidade para os componentes de processamento.[4]
2. **Camada Intermédia (Sinapses Memristivas):** Diretamente sobre a camada CMOS, uma camada de matrizes de barras cruzadas de memristores de HfO_2 1T1R é fabricada utilizando processos compatíveis com o back-end-of-line (BEOL). Esta compatibilidade é crucial, pois significa que a fabricação da camada

de memória pode ocorrer a temperaturas mais baixas que não danificam os transístores CMOS subjacentes.[4] Esta camada funciona como a memória computacional da NCHE v6, onde os pesos sinápticos são armazenados de forma não volátil e as operações de multiplicação de vetor por matriz são executadas na memória.

3. **Camada Superior (Fotónica de Silício):** A camada final é uma camada de fotónica de silício (SiPh) que implementa a Rede Ótica em Chip (ONoC). Esta camada contém todos os componentes óticos necessários, incluindo guias de onda, moduladores e filtros baseados em micro-ressoadores em anel, e acopladores para interface com fibras óticas externas.[5] A sua colocação no topo da pilha facilita o acoplamento de luz para dentro e para fora do chip.

A principal vantagem desta abordagem M3D é a minimização radical da distância física que os sinais devem percorrer entre a computação (neurónios CMOS), a memória (sinapses memristivas) e a comunicação (ONoC fotónica). Ao encurtar os comprimentos das interligações de centímetros (em sistemas multi-chip) para nanómetros ou micrómetros, a latência e a energia associadas ao movimento de dados — o principal fator limitante na computação moderna — são drasticamente reduzidas.[3] Esta integração íntima é o que permite à NCHE v6 superar verdadeiramente o estrangulamento de von Neumann a todos os níveis da arquitetura.

3.2. A Malha de Comunicação: Uma Rede Ótica em Chip (ONoC) Hierárquica

À medida que o número de núcleos neuromórficos num chip aumenta, a comunicação entre eles torna-se o novo estrangulamento de desempenho e energia. As interligações elétricas tradicionais em chip (NoCs) lutam para fornecer a largura de banda necessária em longas distâncias no chip sem consumir uma quantidade proibitiva de energia.[66] Para resolver este problema, a NCHE v6 emprega uma Rede Ótica em Chip (ONoC) como a sua espinha dorsal de comunicação de longo alcance.

A fotónica de silício oferece vantagens transformadoras sobre a eletrónica para a comunicação em chip:

- **Largura de Banda Massiva:** Utilizando a Multiplexagem por Divisão de Comprimento de Onda (WDM), um único guia de onda ótico pode transportar simultaneamente centenas de canais de dados independentes, cada um numa cor (comprimento de onda) de luz diferente. Isto permite larguras de banda

agregadas na ordem dos terabits ou mesmo petabits por segundo.[30]

- **Baixa Latência:** Os sinais óticos propagam-se a uma fração significativa da velocidade da luz, resultando em latências de comunicação que são fundamentalmente mais baixas do que as das interligações elétricas, que são limitadas pela capacitância e resistência dos fios.[66]
- **Eficiência Energética Superior:** Uma vez que um fotão é lançado num guia de onda, ele propaga-se com uma perda muito baixa, eliminando a necessidade de repetidores consumidores de energia que são necessários para sinais elétricos em longas distâncias. Isto resulta numa energia por bit significativamente menor para a comunicação.[67]

A topologia da ONoC da NCHE v6 é concebida para ser escalável e eficiente. Em vez de uma topologia de malha plana, que pode sofrer de congestionamento nos nós centrais, a NCHE v6 utiliza uma **topologia hierárquica baseada em árvore**.[25] Neste esquema, os núcleos neuromórficos são agrupados em clusters. A comunicação dentro de um cluster pode utilizar interligações elétricas curtas e eficientes. No entanto, a comunicação entre clusters é tratada pela ONoC. Cada cluster liga-se a um ramo da árvore ótica, e os dados são encaminhados hierarquicamente através da rede. Esta abordagem é altamente escalável — novos clusters podem ser adicionados como novos ramos — e mantém uma largura de banda local constante, evitando os pontos de estrangulamento de outras topologias.[25]

O mecanismo de comunicação funciona da seguinte forma: quando um neurónio dispara, o seu spike é enviado para um modulador de micro-ressoador em anel. Este dispositivo, sintonizado para um comprimento de onda específico, codifica o spike como um pulso de luz nesse canal de comprimento de onda. O pulso de luz viaja através dos guias de onda da ONoC. Nos clusters de destino, filtros de micro-ressoador em anel, sintonizados para o mesmo comprimento de onda, desviam o pulso de luz do guia de onda principal para um fotodetector local, que converte o sinal ótico de volta num spike elétrico para os neurónios de destino.[5]

3.3. Desempenho Projetado: Eficiência Energética e Densidade de Largura de Banda

Com base no estado da arte da investigação em transceptores de fotónica de silício e ONoCs, a NCHE v6 estabelece metas de desempenho agressivas, mas alcançáveis, para a sua malha de comunicação. Estas métricas quantificam as vantagens da

abordagem M3D-ONoC sobre as arquiteturas convencionais.

- **Eficiência Energética:** A meta para a energia de comunicação é **sub-100 fJ/bit**, com um objetivo ambicioso de **<50 fJ/bit**. Esta projeção é sustentada por demonstrações recentes de transceptores fotónicos integrados que alcançam eficiências na ordem dos 50-70 fJ/bit para o transmissor e o recetor, respetivamente.[72] Isto representa uma melhoria de uma a duas ordens de magnitude em relação às interligações elétricas em chip para distâncias equivalentes.
- **Densidade de Largura de Banda:** A NCHE v6 visa uma **densidade de largura de banda na linha de costa superior a 2 Tbps/mm** e uma **densidade de largura de banda aérea superior a 10 Tbps/mm²**. Estas metas estão alinhadas com as projeções de vários grupos de investigação líderes que trabalham em interligações óticas co-embaladas, que veem estes valores como essenciais para a próxima geração de sistemas de computação.[73] Tal densidade de largura de banda é inatingível com interligações elétricas.
- **Latência:** A latência de comunicação em chip será dominada pelo tempo de voo da luz nos guias de onda de silício, resultando em latências na ordem dos **sub-nanossegundos** para a travessia do chip, permitindo um processamento em tempo real verdadeiramente rápido.[5]

A combinação da integração M3D e de uma ONoC de alto desempenho não é apenas uma melhoria incremental; é uma solução arquitetónica fundamental para o problema do movimento de dados. A computação na memória com memristores resolve o estrangulamento local entre computação e memória.[22] No entanto, para redes em larga escala, a comunicação entre os diferentes bancos de memória computacional torna-se o novo fator limitante.[25] A ONoC quebra este segundo estrangulamento, fornecendo uma espinha dorsal de comunicação com uma largura de banda e eficiência que podem acompanhar a densidade computacional. A M3D é a tecnologia de fabrico que torna possível unir fisicamente estes componentes díspares numa única e poderosa unidade computacional.[3] A tabela seguinte quantifica as vantagens esperadas desta abordagem.

Tabela 3.1: Métricas de Desempenho Projetadas da ONoC da NCHE v6 vs. NoC Elétrica

Métrica	NoC Elétrica (Estado da Arte)	ONoC da NCHE v6 (Projetada)	Vantagem

Densidade de Largura de Banda (Aérea)	~0.1-0.5 Tbps/mm²	> 10 Tbps/mm² [74]	> 20-100x
Eficiência Energética	> 1 pJ/bit para links longos em chip	< 100 fJ/bit, alvo < 50 fJ/bit [72]	> 10-20x
Latência	Dependente do comprimento do fio, >1 ns/mm	Limitada pela velocidade da luz, ~10 ps/mm	> 100x
Escalabilidade com a Distância	A energia/bit aumenta linearmente com a distância.	A energia/bit é largamente independente da distância.	Altamente escalável para chips grandes.
Impacto no Desempenho da Aplicação	O estrangulamento da comunicação limita o desempenho.	Redução do tempo de treino de MLP em 70.12% e da energia em 48.36% em média.[67]	Transformacional

Parte IV: Arquitetura ao Nível da Rede e Codificação da Informação

Para além dos componentes individuais e do substrato físico, a eficácia da NCHE v6 é determinada pela forma como os seus milhões de neurónios e milhares de milhões de sinapses são organizados e pela forma como a informação é codificada e transmitida através deles. A arquitetura da NCHE v6 ao nível da rede inspira-se diretamente na estrutura e nos mecanismos de codificação do cérebro para alcançar uma computação robusta e eficiente. Isto envolve uma topologia de rede específica, esquemas de codificação temporal sofisticados e uma representação de dados formal baseada em grafos.

4.1. Topologia da Rede: Organização Hierárquica de Mundo Pequeno

A conectividade na NCHE v6 não é nem aleatória, nem totalmente conectada, nem

rigidamente em camadas como nas ANNs tradicionais. Em vez disso, implementa uma **topologia de rede de mundo pequeno hierárquica**, uma assinatura da conectividade estrutural e funcional do cérebro humano.[31]

Uma rede de mundo pequeno é caracterizada por duas propriedades estatísticas distintas:

1. **Alto Coeficiente de Agrupamento (C):** Os neurónios estão densamente interligados com os seus vizinhos locais, formando clusters ou módulos altamente conectados. Isto é análogo ao princípio de que "os amigos dos meus amigos são também meus amigos" numa rede social.[31]
2. **Baixo Comprimento Médio do Caminho (L):** Apesar da conectividade predominantemente local, quaisquer dois neurónios na rede podem ser alcançados através de um número surpreendentemente pequeno de passos sinápticos. Isto é conseguido através da existência de "atalhos" esparsos de longo alcance — conexões que ligam clusters distantes.[31]

Esta topologia de mundo pequeno é considerada evolutivamente otimizada para um processamento de informação eficiente. O alto agrupamento suporta a **segregação** de informação, permitindo que módulos neuronais se especializem em computações locais e processamento de características específicas. Os atalhos de longo alcance permitem a **integração** rápida de informação de diferentes módulos, permitindo a ligação global de características e a computação em toda a rede.[32] Esta arquitetura alcança um equilíbrio quase ótimo entre o custo de fiação (minimizando o comprimento total das conexões) e a eficiência da comunicação global.[32] Além disso, as redes de mundo pequeno exibem uma robustez inerente a falhas aleatórias de nós, uma vez que a remoção de um nó não-hub raramente aumenta drasticamente o comprimento médio do caminho.[31]

Na NCHE v6, esta topologia é formada dinamicamente através dos mecanismos de plasticidade estrutural homeostática (HSP). A regra de formação de sinapses dependente da distância favorece a criação de conexões locais (aumentando C), enquanto o impulso homeostático para formar novas conexões quando a atividade está perto do equilíbrio leva à emergência de conexões de longo alcance (diminuindo L).[10]

4.2. Representação da Informação: Codificação Híbrida Temporal e de Taxa

A forma como a informação é representada em picos de atividade neuronal (spikes) é fundamental para a computação em SNNs. A NCHE v6 utiliza um esquema de codificação híbrido e flexível que aproveita a precisão temporal dos spikes para uma computação rápida e eficiente, afastando-se da codificação de taxa simples, que apenas considera a frequência de disparos.[77]

Os esquemas de codificação primários na NCHE v6 são temporais:

1. **Codificação por Tempo até ao Primeiro Spike (TTFS - Time-to-First-Spike):** Para o processamento rápido de estímulos, a intensidade da informação é codificada na latência do primeiro spike emitido por uma população de neurónios. Estímulos mais fortes provocam spikes mais precoces.[78] Este esquema é ideal para ondas de processamento feed-forward rápidas, como as observadas no processamento visual rápido no cérebro, e é altamente eficiente em termos de energia, uma vez que cada neurónio pode precisar de disparar apenas uma vez para transmitir a informação.[78]
2. **Codificação por Ordem de Classificação (ROC - Rank-Order Coding):** Em vez de depender do tempo absoluto de um spike, a ROC codifica a informação na ordem relativa de disparo entre os neurónios de uma população.[81] Por exemplo, para uma população de
   N neurónios, existem N! ordens de disparo possíveis, permitindo uma capacidade de codificação massiva. A ROC é inerentemente mais robusta a ruído e atrasos de transmissão do que a codificação de tempo absoluto e foi proposta como um mecanismo para a transmissão rápida de informação no sistema visual.[81]

Embora a codificação temporal seja a principal, a arquitetura NCHE v6 ainda pode aproveitar a **codificação de taxa**. A taxa de disparo média de um neurónio ou população ao longo de janelas de tempo mais longas serve como um sinal robusto e de baixa resolução. Este sinal é particularmente importante para os mecanismos de plasticidade homeostática, que operam em escalas de tempo mais lentas para regular a atividade da rede [49], e para certas funções de perda durante o treino que podem ser baseadas em taxa.[77] Esta flexibilidade permite que a NCHE v6 utilize o esquema de codificação mais apropriado para a tarefa e a escala de tempo em questão.

4.3. Estruturas de Dados para Redes Esparsas: Uma Abordagem Baseada em Grafos

A representação eficiente da estrutura da rede em software é crucial para a compilação, simulação e otimização. Dado que a topologia de mundo pequeno da NCHE v6 é inerentemente esparsa (ou seja, a maioria das conexões possíveis não existe), representá-la como uma matriz de adjacência densa seria extremamente ineficiente em termos de memória.

Portanto, a NCHE v6 adota formalmente uma representação baseada em grafos. A rede é definida como um grafo direcionado G=(V,E), onde o conjunto de vértices V corresponde aos neurónios e o conjunto de arestas E corresponde às sinapses.[82] Esta não é apenas uma analogia; é uma representação formal que permite a aplicação de um vasto corpo de ferramentas da teoria dos grafos.

Para armazenar e manipular este grafo esparso, a estrutura de dados escolhida é o formato **Compressed Sparse Row (CSR)**, ou a sua variante distribuída (dCSR) para sistemas de múltiplos núcleos.[82] O CSR é um formato padrão da indústria para a representação eficiente de matrizes esparsas. Em vez de armazenar uma matriz

N×N, ele utiliza três vetores para armazenar apenas as entradas não nulas (as sinapses):

- Um vetor de **valores** que armazena os pesos das sinapses.
- Um vetor de **índices de coluna** que armazena a coluna (neurónio de destino) de cada sinapse.
- Um vetor de **ponteiros de linha** que indica onde em o vetor de índices de coluna começam as sinapses para cada linha (neurónio de origem).

Na NCHE v6, esta estrutura de dados CSR é estendida para armazenar não apenas os pesos sinápticos, mas também o rico estado associado a cada vértice (os parâmetros do neurónio de Izhikevich a,b,c,d, o potencial de membrana v, a variável de recuperação u, etc.) e a cada aresta (o peso sináptico w, o traço de elegibilidade eij, os parâmetros de plasticidade, etc.).[82]

Esta abordagem baseada em grafos tem implicações profundas. Trata a rede neuronal não apenas como um "grafo computacional" que descreve uma sequência de operações em camadas, mas como um "grafo relacional" onde a própria estrutura de conectividade é uma parte intrínseca da computação.[83] Estudos mostraram que as propriedades teóricas dos grafos de uma rede, como o seu coeficiente de agrupamento e o comprimento médio do caminho, estão diretamente correlacionadas com o seu desempenho preditivo.[83] Ao adotar uma representação formal de grafo e uma topologia específica (mundo pequeno), a NCHE v6 pode alavancar ferramentas poderosas da teoria dos grafos para a sua otimização, análise e compilação, indo

além do simples mapeamento camada por camada. Isto abre caminho para abordagens mais baseadas em princípios para a pesquisa de arquitetura neuronal e otimização de topologia.

Parte V: Inteligência e Operação ao Nível do Sistema

Uma arquitetura neuromórfica verdadeiramente avançada não é definida apenas pelo seu hardware, mas também pela sua capacidade de ser programada, otimizada e operada de forma inteligente e segura. A NCHE v6 incorpora um conjunto de capacidades ao nível do sistema concebidas para enfrentar os desafios da programação de hardware assíncrono massivamente paralelo, otimizar a sua configuração para um desempenho máximo e garantir um funcionamento robusto e seguro em ambientes do mundo real. Estas capacidades incluem um paradigma de programação de co-design, otimização de inspiração quântica, resiliência bio-inspirada e IA explicável.

5.1. O Paradigma de Programação da NCHE v6: Co-Design de Algoritmo-Hardware

O mapeamento de SNNs complexas para hardware neuromórfico baseado em tiles, com as suas restrições de recursos (por exemplo, um número limitado de sinapses por núcleo), é um desafio significativo.[76] Um mapeamento ingénuo pode resultar numa utilização ineficiente dos recursos, aumento da latência e degradação da precisão. Para resolver este problema, a NCHE v6 adota uma abordagem de programação que vai além de um compilador tradicional, utilizando uma estrutura de

co-exploração de algoritmo-hardware.

Esta abordagem, inspirada em estruturas como a ANCoEF (Asynchronous Neuromorphic Co-Exploration Framework) [11], trata o mapeamento como um problema de otimização multi-objetivo. O processo desenrola-se em várias fases:

1. **Especificação da SNN:** O utilizador define a SNN de alto nível numa linguagem como Python, utilizando uma biblioteca semelhante à Lava da Intel.[38] Esta

especificação descreve a arquitetura da rede, os modelos de neurónios, as regras de plasticidade e os parâmetros iniciais.

2. **Particionamento Consciente do Hardware:** A estrutura analisa o grafo da SNN e particiona-o em clusters de neurónios e sinapses. Este particionamento é "consciente do hardware", o que significa que cada cluster é dimensionado para se ajustar às restrições de recursos de um único núcleo neuromórfico na NCHE v6 (por exemplo, o número máximo de neurónios e sinapses que um núcleo pode suportar).[86]
3. **Otimização Multi-Objetivo:** Esta é a fase mais inovadora. Em vez de usar uma heurística simples para colocar os clusters nos núcleos, a estrutura emprega um agente de **Aprendizagem por Reforço (RL)**. Este agente explora o vasto espaço de design de possíveis mapeamentos (qual cluster vai para qual núcleo) e configurações da ONoC (rotas de comunicação). O objetivo do agente de RL é encontrar uma configuração que otimize simultaneamente múltiplos objetivos: maximizar a precisão da aplicação, minimizar a latência de comunicação e minimizar o consumo total de energia.[11]
4. **Geração de Configuração:** Uma vez que o agente de RL converge para uma solução ótima (ou quase ótima), a estrutura compila esta configuração num ficheiro binário. Este ficheiro é então carregado para o chip NCHE v6 para programar os parâmetros dos neurónios, definir as condutâncias iniciais dos memristores e configurar as tabelas de roteamento da ONoC.

Este paradigma de co-design garante que as aplicações SNN não são apenas "executadas" no hardware, mas são mapeadas de uma forma que explora sinergicamente as suas capacidades arquitetónicas para um desempenho ótimo.

5.2. Otimização de Inspiração Quântica para a Configuração da Rede

A eficácia do motor de aprendizagem em chip da NCHE v6 depende criticamente da sua configuração inicial — a topologia da rede e os hiperparâmetros que governam as regras de plasticidade. Encontrar a configuração ideal é um problema de otimização combinatória NP-difícil, com um espaço de pesquisa demasiado vasto para ser explorado exaustivamente por métodos clássicos.[87]

Para enfrentar este desafio de meta-otimização, a NCHE v6 integra um módulo offline de **Algoritmo Genético de Inspiração Quântica (QIGA)**.[13] Os QIGAs são algoritmos clássicos que se inspiram em conceitos da computação quântica — como qubits e

superposição — para realizar uma exploração mais eficiente e robusta de espaços de soluções complexos do que os algoritmos genéticos tradicionais.[88] Em vez de cromossomas binários, os QIGAs usam "cromossomas de qubits", onde cada gene pode existir numa superposição de 0 e 1, permitindo que um único cromossoma represente simultaneamente múltiplas soluções potenciais.[14]

Na NCHE v6, o QIGA será utilizado para duas tarefas de otimização cruciais, realizadas offline durante a fase de design da rede:

- **Otimização da Topologia da Rede:** O QIGA evoluirá populações de grafos de conectividade para encontrar topologias de mundo pequeno ótimas que são adaptadas a classes específicas de problemas (por exemplo, visão, audição). A função de aptidão avaliará o desempenho da rede (precisão, eficiência) numa tarefa de benchmark.
- **Ajuste de Hiperparâmetros:** O QIGA otimizará o complexo conjunto de hiperparâmetros para o motor de plasticidade unificado (por exemplo, as taxas de aprendizagem e as constantes de tempo para p-STDP, R-STDP e HSP). Isto automatiza um processo que é tipicamente feito através de uma pesquisa em grelha manual, lenta e propensa a erros.

Esta abordagem cria um poderoso ciclo de otimização de dois níveis. No "ciclo interno", a aprendizagem em chip (o motor de plasticidade) adapta a rede aos dados em tempo real. No "ciclo externo", o QIGA otimiza a própria estrutura e as regras de aprendizagem, fornecendo configurações iniciais superiores que permitem que o ciclo interno aprenda de forma mais rápida e eficaz.

5.3. Resiliência e Segurança: Uma Defesa Bio-Inspirada e Multi-Camadas

Os sistemas de computação do mundo real devem ser robustos a falhas e seguros contra ataques. A NCHE v6 aborda a resiliência e a segurança não como características estáticas, mas como propriedades dinâmicas e adaptativas, inspirando-se novamente na biologia.

5.3.1. Auto-Reparação e Tolerância a Falhas via Astrócitos

Os dispositivos em nanoescala, como os memristores, são suscetíveis a falhas. As sinapses podem ficar presas num estado de alta ou baixa resistência (falhas permanentes) ou a sua condutância pode derivar ao longo do tempo (falhas transitórias), o que degrada o desempenho da rede.[92] A abordagem tradicional para a tolerância a falhas envolve redundância de hardware (por exemplo, colunas e linhas sobressalentes), que é estática e dispendiosa.

A NCHE v6 implementa uma solução dinâmica e bio-inspirada: a auto-reparação mediada por astrócitos. A arquitetura inclui circuitos especializados que emulam a função dos astrócitos no cérebro, utilizando o modelo **Leaky Integrate-and-Fire Astrocyte (LIFA)**.[15] Cada circuito de astrócito monitoriza a atividade de uma população local de neurónios. Se um neurónio ou uma sinapse se tornar defeituoso (por exemplo, um neurónio fica silencioso ou uma sinapse fica presa), a mudança na atividade local é detetada pelo astrócito. Em resposta, o astrócito pode iniciar mecanismos de auto-reparação, como:

- Modular a plasticidade das sinapses vizinhas para compensar a conexão defeituosa.
- Desencadear a plasticidade estrutural para podar a conexão defeituosa e promover o crescimento de uma nova para a contornar.

Este mecanismo de reparação autónomo e não supervisionado, inspirado na regulação astrocítica da plasticidade sináptica [95], proporciona uma tolerância a falhas superior a 80%.[15] Trata a resiliência como uma propriedade emergente de um sistema adaptativo, em vez de uma garantia estática.

5.3.2. Segurança a Nível de Hardware

As SNNs, como outras redes neuronais, são vulneráveis a uma série de ataques, incluindo exemplos adversariais (entradas ligeiramente perturbadas que causam classificações erradas) e ataques a nível de hardware, como injeção de falhas (por exemplo, através de falhas de energia) e ataques de canal lateral (extração de informação através do consumo de energia).[97] A NCHE v6 implementa uma estratégia de defesa em várias camadas:

- **Design de Circuito Robusto:** Os circuitos dos neurónios e os seus drivers de corrente são concebidos para serem intrinsecamente mais robustos a flutuações e falhas na fonte de alimentação, mitigando o impacto dos ataques de injeção de

falhas baseados em energia.[99]

- **Deteção de Cavalos de Tróia de Hardware (HT):** Alguns HTs propostos para SNNs dependem de um padrão de picos de entrada específico e invulgar para desencadear uma carga útil maliciosa.[101] A NCHE v6 incluirá monitores em chip que procuram estes padrões de atividade anómalos que não ocorrem em funcionamento normal, sinalizando-os como potenciais gatilhos de HT.
- **Treino Adversarial:** A estrutura de programação da NCHE v6 suporta o treino adversarial. Durante este processo, a rede é explicitamente treinada não só com dados limpos, mas também com exemplos adversariais concebidos para a enganar. Isto força a rede a aprender representações mais robustas, tornando-a menos suscetível a perturbações na entrada.[98]

5.4. IA Explicável (XAI) para Sistemas Neuromórficos

Uma barreira significativa à adoção da IA em domínios críticos como a saúde e as finanças é a natureza de "caixa negra" de muitos modelos. A complexidade dinâmica e não linear das SNNs torna-as particularmente difíceis de interpretar.[34] Para resolver isto, a NCHE v6 integra mecanismos de hardware e software para apoiar a IA Explicável (XAI) adaptada a SNNs.

A abordagem principal é a **Atribuição Temporal de Spikes (TSA - Temporal Spike Attributions)**.[16] Ao contrário dos métodos XAI para ANNs (como LIME ou SHAP) que atribuem importância às características de entrada estáticas (por exemplo, pixels numa imagem), a TSA aproveita a dimensão temporal da computação em SNNs. O mecanismo funciona da seguinte forma:

- Para uma dada decisão de saída (um spike de um neurónio de saída), a TSA rastreia retroativamente a cadeia causal de eventos que levaram a esse disparo.
- Analisa toda a informação disponível nas variáveis internas do modelo: os **tempos dos spikes** de entrada e ocultos, os **pesos sinápticos** que eles atravessaram e os **potenciais de membrana** resultantes nos neurónios da camada de saída.[16]
- O resultado é um mapa de saliência que não destaca apenas *quais* características de entrada foram importantes, mas *quando* os seus spikes correspondentes foram importantes. Isto fornece uma explicação temporalmente precisa do raciocínio da rede.

A arquitetura NCHE v6 fornece "sondas" de hardware, semelhantes às do Loihi [36], que

permitem a leitura em tempo real das variáveis internas necessárias (tempos de spikes, potenciais de membrana) para que o algoritmo TSA construa estas explicações com uma sobrecarga mínima. Isto torna a interpretabilidade uma característica de primeira classe do sistema, em vez de uma análise post-hoc.

Parte VI: NCHE v6 em Aplicação: Benchmarking e Horizontes Futuros

O valor final de qualquer arquitetura de computação reside no seu desempenho em tarefas do mundo real. A NCHE v6 foi concebida não para ser um exercício teórico, mas para ser uma plataforma de alto desempenho que permite uma nova classe de aplicações de IA. Esta secção descreve a estratégia para avaliar o seu desempenho através de benchmarks abrangentes, destaca as suas capacidades avançadas em áreas como a aprendizagem contínua e a robótica, e conclui sobre o seu potencial transformador.

6.1. Avaliação de Desempenho: Uma Estratégia de Benchmarking Abrangente

Avaliar uma arquitetura tão inovadora como a NCHE v6 requer mais do que os benchmarks de IA tradicionais, que são frequentemente dominados por tarefas de classificação de imagens estáticas em grandes lotes e não conseguem capturar os pontos fortes dos sistemas neuromórficos, como a eficiência em tarefas temporais e esparsas.[104] Portanto, a NCHE v6 será avaliada utilizando um conjunto de suites de benchmarks neuromórficos, concebidos pela comunidade para fornecer uma avaliação justa e representativa.

As principais suites de benchmarks a serem utilizadas incluem:

1. **NeuroBench:** Uma suite de benchmarks colaborativa e impulsionada pela comunidade, com contribuições da academia e da indústria. A NeuroBench fornece um conjunto de tarefas e métricas padrão para comparar de forma justa soluções neuromórficas entre si e com abordagens não neuromórficas. Inclui uma "faixa de algoritmo" independente do hardware e uma "faixa de sistema" dependente do hardware, abrangendo domínios de aplicação como visão e

audição.[105]

2. **SNABSuite (Spiking-Neural-Network-Application-Benchmark-Suite):** Uma suite de benchmarks abrangente que se concentra na comparação entre diferentes plataformas de hardware neuromórfico, incluindo sistemas digitais e de sinal misto. A SNABSuite abrange desde a caracterização de baixo nível até à avaliação de aplicações de alto nível e, crucialmente, inclui um modelo de energia que permite estimar o consumo de energia de uma rede num sistema alvo sem ter acesso direto a ele.[107]
3. **NeuroSeqBench:** Uma suite de benchmarks especializada, concebida para preencher uma lacuna crítica nas avaliações existentes: a capacidade de processamento temporal. A NeuroSeqBench inclui tarefas com dinâmicas temporais ricas e dependências de longo alcance, que são precisamente o tipo de problemas para os quais a dinâmica rica dos neurónios de Izhikevich e a arquitetura recorrente da NCHE v6 são mais adequadas.[110]

A avaliação nestes benchmarks irá além da simples precisão de classificação para incluir um conjunto holístico de **Métricas de Desempenho Chave (KPIs)**:

- **Eficiência Energética:** Medida em energia por inferência (Joules/inferência) e o produto energia-atraso (EDP).
- **Latência:** Tempo até à primeira resposta (tempo até ao primeiro spike) e tempo total de processamento.
- **Densidade Computacional:** Operações sinápticas por segundo por watt (SOPS/W), uma medida da produção computacional por unidade de energia.
- **Desempenho de Aprendizagem:** Em tarefas de aprendizagem contínua, serão medidas métricas como a velocidade de adaptação a novas tarefas e a taxa de esquecimento de tarefas antigas.

A seleção destes benchmarks e métricas reflete o foco estratégico da NCHE v6. O objetivo não é necessariamente superar uma GPU na classificação do ImageNet, mas sim demonstrar uma superioridade de ordens de magnitude em tarefas dinâmicas, temporais e com restrições de energia, onde a IA atual falha.

Tabela 6.1: Metas de Desempenho da NCHE v6 em Tarefas Chave da NeuroBench

Tarefa de Benchmark (NeuroBench)	Métrica Principal	Linha de Base (SotA - ex: Loihi 2)	Meta de Desempenho da NCHE v6	Justificação da Melhoria

Reconhecimento de Palavras-Chave (Spiking)	Precisão / Energia por Inferência	~90% / ~50 µJ [105]	> 95% / < 5 µJ	A plasticidade multi-fator e os memristores analógicos permitem uma aprendizagem mais fina e uma computação mais eficiente.
Reconhecimento de Gestos (Dados DVS)	Precisão / Latência	~85% / ~10 ms [106]	> 90% / < 1 ms	A ONoC e o processamento assíncrono de ponta a ponta reduzem drasticamente a latência de comunicação.
Classificação de Áudio (Spiking Heidelberg Digits)	Precisão	~80% [105]	> 90%	Os neurónios de Izhikevich e a codificação temporal rica capturam melhor as características temporais complexas dos dados de áudio.
Tarefa de Aprendizagem Contínua	Taxa de Esquecimento	Alta (sem mecanismos CL)	< 5%	A plasticidade homeostática e estrutural (HSP) mitiga ativamente o esquecimento catastrófico.[18]

6.2. Capacidades Avançadas: Permitindo a Aprendizagem Contínua e a Robótica Autónoma

As características arquitetónicas da NCHE v6 convergem para permitir capacidades

que estão na vanguarda da investigação em IA.

Aprendizagem Contínua: O cérebro humano aprende continuamente ao longo da vida, adaptando-se a novas informações sem apagar abruptamente as memórias antigas. A maioria dos sistemas de IA sofre de "esquecimento catastrófico" quando treinada sequencialmente em novas tarefas.[18] A NCHE v6 foi concebida para resolver este "dilema estabilidade-plasticidade". A sua plasticidade Hebbiana (p-STDP) e orientada por objetivos (R-STDP) fornece a

plasticidade necessária para adquirir novos conhecimentos. Ao mesmo tempo, os seus mecanismos de plasticidade homeostática e estrutural (HSP) fornecem a **estabilidade**, regulando a atividade da rede e consolidando as memórias através da reorganização estrutural.[18] Esta interação dinâmica permite que a NCHE v6 aprenda continuamente a partir de um fluxo de dados, tornando-a uma plataforma ideal para sistemas de IA ao longo da vida.

Robótica Autónoma e Navegação: As plataformas robóticas, como drones e robôs móveis, são severamente limitadas por restrições de Tamanho, Peso e Potência (SWaP).[26] Executar modelos de IA complexos nestas plataformas é um desafio significativo. A NCHE v6 aborda diretamente este problema. O seu consumo de energia ultra-baixo, o processamento em tempo real de sensores orientados a eventos (como câmaras DVS) e a capacidade de aprendizagem em chip são críticos para a IA de ponta.[27] Um sistema NCHE v6 num robô poderia realizar tarefas como o reconhecimento de lugar visual (localização) e a navegação em tempo real, consumindo menos de 8% da energia exigida pelos métodos convencionais.[26] A capacidade de aprender e adaptar-se online (por exemplo, aprender a navegar num novo ambiente) sem depender de um data center na nuvem é transformadora para a autonomia robótica.[28]

6.3. Conclusão: A NCHE v6 como um Catalisador para a Próxima Geração de IA

A arquitetura de Computação Neuromórfica de Alta Eficiência, Versão 6, não é uma melhoria incremental das tecnologias existentes. É um blueprint para uma nova classe de sistemas de computação, concebidos em torno dos princípios fundamentais da computação biológica. Ao integrar holisticamente inovações em modelos de neurónios (Izhikevich), dispositivos sinápticos (memristores de HfO_2), comunicação em chip (ONoC fotónica), paradigmas de aprendizagem (plasticidade multi-fator) e

inteligência ao nível do sistema (co-design, QIO, auto-reparação, XAI), a NCHE v6 oferece um caminho viável para superar as limitações fundamentais que restringem a IA atual.

O resultado é uma arquitetura que promete não apenas ser ordens de magnitude mais rápida e eficiente em termos de energia, mas também fundamentalmente mais inteligente. A NCHE v6 foi concebida para ser adaptativa, aprendendo continuamente com a sua experiência; robusta, tolerando falhas de hardware e resistindo a ataques; e transparente, fornecendo explicações para as suas decisões. Ao fornecer este blueprint abrangente e rigorosamente fundamentado, pretendemos catalisar o desenvolvimento da próxima geração de inteligência artificial — uma IA que pode finalmente sair do data center e operar de forma eficiente, autónoma e segura nas complexidades do mundo real.

Works cited

1. Improving the Izhikevich Model Based on Rat Basolateral Amygdala ..., accessed July 3, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC7253815/
2. Izhikevich-Inspired Temporal Dynamics for Enhancing Privacy, Efficiency, and Transferability in Spiking Neural Networks - arXiv, accessed July 3, 2025, https://arxiv.org/html/2505.04034v1
3. Formation techniques for upper active channel in monolithic 3D integration: an overview, accessed July 3, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10825103/
4. Monolithic three-dimensional integration of RRAM-based hybrid memory architecture for one-shot learning - PMC, accessed July 3, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10628152/
5. Silicon Photonics for Neuromorphic Computing and Artificial Intelligence | Request PDF, accessed July 3, 2025, https://www.researchgate.net/publication/357943747_Silicon_Photonics_for_Neuromorphic_Computing_and_Artificial_Intelligence
6. Multiplicative Spike-Time-Dependent Plasticity with Metal Oxide Memristors - arXiv, accessed July 3, 2025, https://arxiv.org/pdf/1505.05549
7. Probabilistic metaplasticity for continual learning with memristors in spiking networks - arXiv, accessed July 3, 2025, https://arxiv.org/pdf/2403.08718
8. Three-Factor Learning in Spiking Neural Networks: An Overview of Methods and Trends from a Machine Learning Perspective - arXiv, accessed July 3, 2025, https://arxiv.org/html/2504.05341v1
9. Incorporating structural plasticity into self-organization recurrent networks for sequence learning - Frontiers, accessed July 3, 2025, https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2023.1224752/full
10. Homeostatic structural plasticity – a key to neuronal network formation and

repair - PMC, accessed July 3, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC4125105/
11. ANCoEF: Asynchronous Neuromorphic Algorithm/Hardware Co-Exploration Framework with a Fully Asynchronous Simulator - arXiv, accessed July 3, 2025, http://www.arxiv.org/pdf/2411.06059
12. AI Hardware-Algorithm Co-Design | NanoX Lab - Purdue College of Engineering, accessed July 3, 2025, https://engineering.purdue.edu/NanoX/projects/codesign/
13. Quantum-Inspired Algorithms for AI and Machine Learning - ResearchGate, accessed July 3, 2025, https://www.researchgate.net/publication/385350757_Quantum-Inspired_Algorithms_for_AI_and_Machine_Learning
14. (PDF) QGAIC: Quantum Inspired Genetic Algorithm for Image Classification - ResearchGate, accessed July 3, 2025, https://www.researchgate.net/publication/388231870_QGAIC_Quantum_Inspired_Genetic_Algorithm_for_Image_Classification
15. Neuromorphic Circuits with Spiking Astrocytes for Increased Energy Efficiency, Fault Tolerance, and Memory Capacitance - arXiv, accessed July 3, 2025, https://arxiv.org/html/2502.20492v1
16. Feature Attribution Explanations for Spiking Neural Networks - Bohrium, accessed July 3, 2025, https://www.bohrium.com/paper-details/feature-attribution-explanations-for-spiking-neural-networks/928713876365640035-108619
17. Energy Efficiency of Neuromorphic Hardware Practically Proven - Human Brain Project, accessed July 3, 2025, https://www.humanbrainproject.eu/en/follow-hbp/news/2022/05/24/energy-efficiency-neuromorphic-hardware-practically-proven/
18. Continual Learning with Neuromorphic Computing: Theories, Methods, and Applications, accessed July 3, 2025, https://arxiv.org/html/2410.09218v2
19. Design of CMOS-memristor hybrid synapse and its application for noise-tolerant memristive spiking neural network - Frontiers, accessed July 3, 2025, https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2025.1516971/full
20. Spiking Neural Network Architectures | by NeuroCortex.AI - Medium, accessed July 3, 2025, https://medium.com/@theagipodcast/spiking-neural-network-architectures-e6983ff481c2
21. A Scatter-and-Gather Spiking Convolutional Neural Network on a Reconfigurable Neuromorphic Hardware - Bohrium, accessed July 3, 2025, https://www.bohrium.com/paper-details/a-scatter-and-gather-spiking-convolutional-neural-network-on-a-reconfigurable-neuromorphic-hardware/813236091128643585-11117
22. New Memristor-Based Crossbar Array Architecture with 50-% Area Reduction and 48-% Power Saving for Matrix-Vector Multiplication of Analog Neuromorphic Computing | Request PDF - ResearchGate, accessed July 3, 2025, https://www.researchgate.net/publication/271057848_New_Memristor-Based_Cr

ossbar_Array_Architecture_with_50-_Area_Reduction_and_48-_Power_Saving_for_Matrix-Vector_Multiplication_of_Analog_Neuromorphic_Computing
23. (PDF) Low-Power Memristor for Neuromorphic Computing: From Materials to Applications, accessed July 3, 2025, https://www.researchgate.net/publication/390763975_Low-Power_Memristor_for_Neuromorphic_Computing_From_Materials_to_Applications
24. Neomorphic Quantum computing a possible solution to GPU's and AI on chain - General, accessed July 3, 2025, https://forum.dfinity.org/t/neomorphic-quantum-computing-a-possible-solution-to-gpu-s-and-ai-on-chain/37241
25. Low-latency hierarchical routing of reconfigurable neuromorphic systems - Frontiers, accessed July 3, 2025, https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2025.1493623/full
26. A compact neuromorphic system for ultra-energy-efficient, on-device robot localization, accessed July 3, 2025, https://arxiv.org/html/2408.16754v2
27. Neuromorphic computing hardware and neural architectures for robotics - Yulia Sandamirskaya, accessed July 3, 2025, https://sandamirskaya.eu/resources/SandamirskayaEtAl2022_SciRob.pdf
28. [2503.09636] Real-Time Neuromorphic Navigation: Guiding Physical Robots with Event-Based Sensing and Task-Specific Reconfigurable Autonomy Stack - arXiv, accessed July 3, 2025, https://arxiv.org/abs/2503.09636
29. A review on monolithic 3D integration: From bulk semiconductors to low-dimensional materials - SciOpen, accessed July 3, 2025, https://www.sciopen.com/article/10.26599/NR.2025.94907225
30. High-bandwidth density silicon photonic resonators for energy-efficient optical interconnects | Applied Physics Reviews | AIP Publishing, accessed July 3, 2025, https://pubs.aip.org/aip/apr/article/10/4/041306/2921480/High-bandwidth-density-silicon-photonic-resonators
31. Small-world network - Wikipedia, accessed July 3, 2025, https://en.wikipedia.org/wiki/Small-world_network
32. Small-world human brain networks: Perspectives and challenges - Helab@BNU, accessed July 3, 2025, https://helab.bnu.edu.cn/wp-content/uploads/pdf/Liao_NBR2017.pdf
33. Neuromorphic algorithms for brain implants: a review - Frontiers, accessed July 3, 2025, https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2025.1570104/full
34. Explainable AI-empowered Neuromorphic Computing Framework for Consumer Healthcare, accessed July 3, 2025, https://www.researchgate.net/publication/382907425_Explainable_AI-empowered_Neuromorphic_Computing_Framework_for_Consumer_Healthcare
35. Event2Vec: Processing neuromorphic events directly by representations in vector space, accessed July 3, 2025, https://arxiv.org/html/2504.15371v1
36. Event-based attention and tracking on neuromorphic hardware - Robotics and

Perception Group, accessed July 3, 2025, https://rpg.ifi.uzh.ch/CVPR19_event_vision_workshop_files/docs/2019CVPRW_Event-based_attention_and_tracking_on_neuromorphic_hardware.pdf
37. SC-IZ: A Low-Cost Biologically Plausible Izhikevich Neuron for Large-Scale Neuromorphic Systems Using Stochastic Computing - MDPI, accessed July 3, 2025, https://www.mdpi.com/2079-9292/13/5/909
38. A Look at Loihi 2 - Intel - Open Neuromorphic, accessed July 3, 2025, https://open-neuromorphic.org/neuromorphic-computing/hardware/loihi-2-intel/
39. Forming-less flexible memristor crossbar array for neuromorphic computing applications produced using low-temperature atomic layer deposition | Request PDF - ResearchGate, accessed July 3, 2025, https://www.researchgate.net/publication/381074907_Forming-less_flexible_memristor_crossbar_array_for_neuromorphic_computing_applications_produced_using_low-temperature_atomic_layer_deposition
40. Flexible HfO2-based ferroelectric memristor | Request PDF - ResearchGate, accessed July 3, 2025, https://www.researchgate.net/publication/363302772_Flexible_HfO2-based_ferroelectric_memristor
41. Y-Doped HfO2 Ferroelectric Memristor for Information Processing and Neuromorphic Computing | ACS Applied Materials & Interfaces - ACS Publications, accessed July 3, 2025, https://pubs.acs.org/doi/abs/10.1021/acsami.5c05846
42. Ultralow Powered 2D MoS2-Based Memristive Crossbar Array for ..., accessed July 3, 2025, https://pubs.acs.org/doi/10.1021/acsami.5c00688
43. Y-Doped HfO2 Ferroelectric Memristor for Information Processing and Neuromorphic Computing - PubMed, accessed July 3, 2025, https://pubmed.ncbi.nlm.nih.gov/40407278/
44. [2405.10909] Memristive response and neuromorphic functionality of polycrystalline ferroelectric Ca:HfO$_{2}$-based devices - arXiv, accessed July 3, 2025, https://arxiv.org/abs/2405.10909
45. Current Opinions on Memristor-Accelerated Machine Learning Hardware - arXiv, accessed July 3, 2025, https://arxiv.org/html/2501.12644v1
46. An Adaptive STDP Learning Rule for Neuromorphic Systems - PMC, accessed July 3, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC8498208/
47. Multi-layer network utilizing rewarded spike time dependent plasticity to learn a foraging task, accessed July 3, 2025, https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005705
48. A Spiking Network Model of Decision Making Employing Rewarded STDP | PLOS One, accessed July 3, 2025, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0090821
49. Homeostatic plasticity - Wikipedia, accessed July 3, 2025, https://en.wikipedia.org/wiki/Homeostatic_plasticity
50. Spike-Timing Dependence of Structural Plasticity Explains Cooperative Synapse Formation in the Neocortex | PLOS Computational Biology, accessed July 3, 2025, https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002689

51. The Reward-Modulated Self-Organizing Recurrent Neural Network... - ResearchGate, accessed July 3, 2025, https://www.researchgate.net/figure/The-Reward-Modulated-Self-Organizing-Recurrent-Neural-Network-RM-SORN-Excitatory-units_fig7_274728362
52. [2303.08530] Combined effects of STDP and homeostatic structural plasticity on coherence resonance - arXiv, accessed July 3, 2025, https://arxiv.org/abs/2303.08530
53. A compound memristive synapse model for statistical learning through STDP in spiking neural networks - PMC - PubMed Central, accessed July 3, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC4267210/
54. A Physics-based Model of RRAM Probabilistic Switching for Generating Stable and Accurate Stochastic Bit-streams | Request PDF - ResearchGate, accessed July 3, 2025, https://www.researchgate.net/publication/339262183_A_Physics-based_Model_of_RRAM_Probabilistic_Switching_for_Generating_Stable_and_Accurate_Stochastic_Bit-streams
55. Modulation of Spike-Timing Dependent Plasticity ... - Frontiers, accessed July 3, 2025, https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2018.00049/full
56. Learning to learn online with neuromodulated synaptic plasticity in spiking neural networks, accessed July 3, 2025, https://www.biorxiv.org/content/10.1101/2022.06.24.497562.full
57. arXiv:2109.05539v5 [cs.NE] 7 Jul 2022, accessed July 3, 2025, https://arxiv.org/pdf/2109.05539
58. First-spike based visual categorization using reward-modulated STDP - CerCo, accessed July 3, 2025, https://cerco.cnrs.fr/wp-content/uploads/2020/02/1705.09132.pdf
59. Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity - BSTU Laboratory of Artificial Neural Networks, accessed July 3, 2025, https://neuro.bstu.by/ai/Turkey-collabolation/06_modulated_STDP.pdf
60. The interplay between homeostatic synaptic scaling and homeostatic structural plasticity maintains the robust firing rate of neural networks - eLife, accessed July 3, 2025, https://elifesciences.org/reviewed-preprints/88376
61. Activity-dependent structural plasticity - PubMed, accessed July 3, 2025, https://pubmed.ncbi.nlm.nih.gov/19162072/
62. What is Synaptic Pruning? - News-Medical, accessed July 3, 2025, https://www.news-medical.net/health/What-is-Synaptic-Pruning.aspx
63. Mechanisms governing activity-dependent synaptic pruning in the mammalian CNS - PMC, accessed July 3, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC8541743/
64. Homeostatic structural plasticity increases the efficiency of small-world networks - Frontiers, accessed July 3, 2025, https://www.frontiersin.org/journals/synaptic-neuroscience/articles/10.3389/fnsyn.2014.00007/full

65. Monolithic silicon-photonic platforms in state-of-the-art CMOS SOI processes [Invited], accessed July 3, 2025, https://sclaser.mit.edu/wp-content/uploads/2018/11/PubOpticsExpress_May07_2018.pdf
66. Optical Network on Chip: A Comprehensive Guide - Number Analytics, accessed July 3, 2025, https://www.numberanalytics.com/blog/optical-network-on-chip-ultimate-guide
67. Optical Versus Electrical: Performance Evaluation of Network On-Chip Topologies for UWASN Manycore Processors | Request PDF - ResearchGate, accessed July 3, 2025, https://www.researchgate.net/publication/334860802_Optical_Versus_Electrical_Performance_Evaluation_of_Network_On-Chip_Topologies_for_UWASN_Manycore_Processors
68. Silicon Photonics for Neuromorphic Computing: A New Era - Number Analytics, accessed July 3, 2025, https://www.numberanalytics.com/blog/silicon-photonics-neuromorphic-computing-era
69. Photonic Breakthroughs in Chip Design: A Survey of Optical Router Architectures, accessed July 3, 2025, https://communities.springernature.com/posts/photonic-breakthroughs-in-chip-design-a-survey-of-optical-router-architectures
70. Unlocking Neuromorphic Computing with Silicon Photonics, accessed July 3, 2025, https://www.numberanalytics.com/blog/silicon-photonics-neuromorphic-computing-guide
71. High-bandwidth density silicon photonic resonators for energy-efficient optical interconnects - Content Delivery Network (CDN), accessed July 3, 2025, https://bpb-us-e1.wpmucdn.com/sites.dartmouth.edu/dist/c/2572/files/2025/01/High-bandwidth_density_silicon_photonic_resonators_for_energy-efficient_optical_interconnects.pdf
72. Hybrid 14nm FinFET - Silicon Photonics Technology for Low-Power Tb/s/mm 2 Optical I/O, accessed July 3, 2025, https://www.researchgate.net/publication/328983930_Hybrid_14nm_FinFET_-_Silicon_Photonics_Technology_for_Low-Power_Tbsmm_2_Optical_IO
73. Silicon Photonics Chip I/O for Ultra High-Bandwidth and Energy-Efficient Die-to-Die Connectivity, accessed July 3, 2025, https://lightwave.ee.columbia.edu/sites/default/files/content/publications/2024/IEEE_CICC_2024.pdf
74. Co-Designed Silicon Photonics Chip I/O for Energy-Efficient Petascale Connectivity, accessed July 3, 2025, https://lightwave.ee.columbia.edu/sites/default/files/content/docs/Papers/2024/Co-Designed_Silicon_Photonics_Chip_I_O_for_Energy-Efficient_Petascale_Connectivity.pdf
75. Exploring the Use of Photonics in Neuromorphic Computing - AZoOptics, accessed July 3, 2025, https://www.azooptics.com/Article.aspx?ArticleID=2753

76. Compiling Spiking Neural Networks to Mitigate Neuromorphic Hardware Constraints, accessed July 3, 2025, https://par.nsf.gov/servlets/purl/10295485
77. Exploring Loss Functions for Time-based Training Strategy in Spiking Neural Networks, accessed July 3, 2025, https://papers.neurips.cc/paper_files/paper/2023/file/cde874a797a8300da693d5e412b7fdc0-Paper-Conference.pdf
78. Analyzing time-to-first-spike coding schemes: A theoretical approach - PMC, accessed July 3, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC9548614/
79. Analyzing time-to-first-spike coding schemes: A theoretical approach - Frontiers, accessed July 3, 2025, https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2022.971937/full
80. Comparison of discriminability for the three coding schemes during... - ResearchGate, accessed July 3, 2025, https://www.researchgate.net/figure/Comparison-of-discriminability-for-the-three-coding-schemes-during-propagation-The_fig4_363846375
81. Inherent trade-off in noisy neural communication with rank-order coding | Phys. Rev. Research - Physical Review Link Manager, accessed July 3, 2025, https://link.aps.org/doi/10.1103/PhysRevResearch.6.L012009
82. Distributed Compressed Sparse Row Format for Spiking Neural Network Simulation, Serialization, and Interoperability - OSTI, accessed July 3, 2025, https://www.osti.gov/servlets/purl/2432223
83. Graph Structure of Neural Networks - Stanford Computer Science, accessed July 3, 2025, https://www-cs.stanford.edu/~jure/pubs/nn_structure-icml20.pdf
84. (PDF) Compiling Spiking Neural Networks to Mitigate Neuromorphic, accessed July 3, 2025, https://www.researchgate.net/publication/346510506_Compiling_Spiking_Neural_Networks_to_Mitigate_Neuromorphic_Hardware_Constraints
85. Neuromorphic Computing and Engineering with AI | Intel®, accessed July 3, 2025, https://www.intel.com/content/www/us/en/research/neuromorphic-computing.html
86. Compiling Spiking Neural Networks to Mitigate Neuromorphic Hardware Constraints | Request PDF - ResearchGate, accessed July 3, 2025, https://www.researchgate.net/publication/350857189_Compiling_Spiking_Neural_Networks_to_Mitigate_Neuromorphic_Hardware_Constraints
87. Incorporating Structural Plasticity Approaches in Spiking Neural Networks for EEG Modelling, accessed July 3, 2025, https://www.researchgate.net/publication/353476135_Incorporating_Structural_Plasticity_Approaches_in_Spiking_Neural_Networks_for_EEG_Modelling
88. Quantum-Inspired Algorithms - GeeksforGeeks, accessed July 3, 2025, https://www.geeksforgeeks.org/artificial-intelligence/quantum-inspired-algorithms/
89. Quantum-Inspired Algorithms: A Comprehensive Guide - Number Analytics, accessed July 3, 2025, https://www.numberanalytics.com/blog/quantum-inspired-algorithms-guide

90. Quantum-Inspired Genetic Algorithms for Combinatorial Optimization Problems - Hasmed, accessed July 3, 2025, https://hasmed.org/index.php/jourasy/article/download/47/66/465
91. Quantum Inspired Genetic Algorithm - Knowledge Engineering and Discovery Research Institute - AUT, accessed July 3, 2025, https://www.aut.ac.nz/kedri-old-site/R-and-D-Systems/quantum-inspired-genetic-algorithm
92. A Design Methodology for Fault-Tolerant Neuromorphic Computing Using Bayesian Neural Network - MDPI, accessed July 3, 2025, https://www.mdpi.com/2072-666X/14/10/1840
93. Fault tolerance in neuromorphic computing systems | Request PDF - ResearchGate, accessed July 3, 2025, https://www.researchgate.net/publication/330487844_Fault_tolerance_in_neuromorphic_computing_systems
94. Fault tolerance in memristive crossbar-based neuromorphic computing systems - CUHK CSE, accessed July 3, 2025, https://www.cse.cuhk.edu.hk/~byu/papers/J45-JVLSI2020-FT-NCS.pdf
95. Astromorphic Self-Repair of Neuromorphic Hardware Systems, accessed July 3, 2025, https://ojs.aaai.org/index.php/AAAI/article/view/25947/25719
96. [2209.07428] Astromorphic Self-Repair of Neuromorphic Hardware Systems - arXiv, accessed July 3, 2025, https://arxiv.org/abs/2209.07428
97. ADVERSARIAL ATTACKS ON SPIKING CONVOLU- TIONAL NETWORKS FOR EVENT-BASED VISION - OpenReview, accessed July 3, 2025, https://openreview.net/pdf?id=e0uknAgETh
98. Adversarial Training for Probabilistic Spiking Neural Networks | Request PDF, accessed July 3, 2025, https://www.researchgate.net/publication/327264961_Adversarial_Training_for_Probabilistic_Spiking_Neural_Networks
99. Fault Injection Attacks in Spiking Neural Networks and Countermeasures - Frontiers, accessed July 3, 2025, https://www.frontiersin.org/journals/nanotechnology/articles/10.3389/fnano.2021.801999/full
100. Fault Injection Attacks in Spiking Neural Networks and Countermeasures, accessed July 3, 2025, https://par.nsf.gov/biblio/10352468-fault-injection-attacks-spiking-neural-networks-countermeasures
101. Input-Triggered Hardware Trojan Attack on Spiking Neural NetworksThis work was supported by the French National Research Agency (ANR) and the UK Research and Innovation (UKRI), Engineering and Physical Sciences Research Council (EPSRC), through the European CHIST-ERA program under the project TruBrain (Grants N^"o" ANR-23 - arXiv, accessed July 3, 2025, https://arxiv.org/html/2503.21793v1
102. Xai Explainable Ai - Lark, accessed July 3, 2025, https://www.larksuite.com/en_us/topics/ai-glossary/xai-explainable-ai
103. Gradient-based feature-attribution explainability methods for spiking neural

networks, accessed July 3, 2025, https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2023.1153999/full
104. Energy Aware Development of Neuromorphic Implantables: From Metrics to Action - arXiv, accessed July 3, 2025, https://arxiv.org/html/2506.09599v1
105. NeuroBench: Advancing Neuromorphic Computing through Collaborative, Fair and Representative Benchmarking, accessed July 3, 2025, https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=936693
106. [2304.04640] NeuroBench: A Framework for Benchmarking Neuromorphic Computing Algorithms and Systems - arXiv, accessed July 3, 2025, https://arxiv.org/abs/2304.04640
107. Benchmarking Neuromorphic Hardware and Its Energy Expenditure - PubMed, accessed July 3, 2025, https://pubmed.ncbi.nlm.nih.gov/35720731/
108. Benchmarking Neuromorphic Hardware and Its Energy Expenditure - Frontiers, accessed July 3, 2025, https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2022.873935/full
109. Benchmarking Neuromorphic Hardware and Its Energy Expenditure - PMC, accessed July 3, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC9201569/
110. liyc5929/neuroseqbench: A benchmark suite for evaluating Spiking Neural Networks (SNNs) on temporal processing tasks, comparing abilities of SNN-related models and learning algorithms for extended temporal sequences. - GitHub, accessed July 3, 2025, https://github.com/liyc5929/neuroseqbench
111. Brain-inspired global-local learning incorporated with neuromorphic computing - arXiv, accessed July 3, 2025, https://arxiv.org/abs/2006.03226
112. Real-Time Neuromorphic Navigation: Guiding Physical Robots with Event-Based Sensing and Task-Specific Reconfigurable Autonomy Stack - arXiv, accessed July 3, 2025, https://arxiv.org/html/2503.09636v1