计算机研究与发展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(11): 2663−2675, 2025

基于混合模式路由的脑启发片上网络架构

王智超[1,2] 陈 亮[1] 李千鹏[1,2] 陈奥新[1,2] 刘 昕[1] 宋文娜[1]
[1](中国科学院自动化研究所 北京 100190)
[2](中国科学院大学人工智能学院 北京 100049)
(wangzhichao2022@ia.ac.cn)

A Brain-Inspired Network-on-Chip Architecture with Hybrid-Mode Routing

Wang Zhichao[1,2], Chen Liang[1], Li Qianpeng[1,2], Chen Aoxin[1,2], Liu Xin[1], and Song Wenna[1]
[1](*Institute of Automation*, *Chinese Academy of Sciences*, *Beijing* 100190)
[2](*School of Artificial Intelligence*, *University of Chinese Academy of Sciences*, *Beijing* 100049)

Abstract The rise and development of large-scale neuromorphic platforms require network-on-chip to support efficient data transmission mechanisms. Although many efforts have been made to develop high performance topology architectures and routing schemes, they still suffer from single transmission mode or poor scalability, making them stay on a low efficiency in neuromorphic computing. Inspired by the small-world properties of human brain networks, this brief proposes an efficient region-broadcast (ReB) routing scheme to support unicast, multicast, and broadcast transmission modes. Besides, a synaptic connections indexing method is deployed to accommodate the ReB routing scheme and support this hybrid-mode packet transmission. This method replaces the traditional multicast routing table, effectively improving network scalability and reducing power consumption. Experimental results show that compared with existing work, the ReB routing scheme reduces the peak spike traffic and link load standard deviation by 11.5% and 20.4%, respectively. The ReB routing scheme brings improvements in latency, throughput, and energy under the validation of synthetic traffic, spiking neural network applications and brain cortical networks. Various synthetic traffic patterns are used in the experiments. The datasets used in spiking neural network applications include MNIST, QTDB, Ev-object, and DVS-Gesture. Finally, the proposed ReB router has an excellent bandwidth of 0.24 spike/cycle and only consumes an area of 0.014 mm^2.

Key words network-on-chip (NoC); brain-inspired; routing scheme; hybrid-mode router; synaptic connections indexing

摘 要 大规模神经形态平台的兴起和发展要求片上网络(network-on-chip, NoC)具备高效的数据传输机制. 现有工作在开发高性能路由拓扑架构和设计路由策略方面已经做出了许多努力,但它们仍然受限于单一传输模式或扩展性差的问题,这导致神经形态计算的效率低. 受人脑网络小世界特性的启发,提出了一种高效的片上网络路由方案——区域广播(region-broadcast, ReB),能够直接支持单播、多播和广播的混合传输模式. 此外,部署了一种突触连接索引方法,以适应所提出的路由方案并支持这种混合模式的传输. 这种方法替代了传统的多播路由表,有效提高了网络扩展性并降低功耗. 实验结果表明,与现有工作相比,ReB路由方案将峰值脉冲流量和链路负载标准差分别降低了11.5%和20.4%. 在合成流量、脉冲神经网络应用和脑皮质柱网络验证下,ReB策略有效提升了片上网络的延迟、吞吐量和功耗等方面的性能. 最后,所提出的ReB路由器的带宽达到0.24 spike/cycle,硬件实现面积仅为0.014 mm^2.

收稿日期: 2024-08-13; **修回日期**: 2025-03-19
基金项目: 科技创新2030-"脑科学与类脑研究"重大项目(2021ZD0200300)
This work was supported by the National Key Research and Development Program of China (2021ZD0200300).
通信作者: 陈亮(liang.chen@ia.ac.cn)

关键词 片上网络;脑启发;路由方案;混合模式路由器;突触连接索引

中图法分类号 TP389.1

DOI: 10.7544/issn1000-1239.202440683 CSTR: 32373.14.issn1000-1239.202440683

目前,人工智能和深度学习发展迅猛,取得了显著的技术突破和应用成果.深度学习算法在图像识别、自然语言处理、语音识别和自动驾驶等领域表现出卓越的性能,推动了智能产品和服务的普及.尽管取得了巨大成就,人工智能和深度学习仍面临挑战,包括对大数据和高性能计算资源的依赖、能效和并行处理能力的限制,以及在复杂任务中的泛化能力不足等.人类大脑由大约 1 000 亿个神经元组成,每个神经元平均有 7 000 个突触,形成了一个错综复杂的巨大网络[1-2].尽管中枢神经系统内的神经元之间有着复杂的联系,大脑仍然通过其高效的事件驱动通信实现了显著的能量效率.可见,大脑和深度学习架构之间存在着巨大的能效差距和并行处理能力的差距[3].

许多国家已经开始从人类大脑的连接模式和计算模型中寻求灵感,通过探索生物神经网络的机制,模拟大脑处理和存储信息的方式,旨在设计受大脑启发的高能效计算平台——神经形态平台(neuromorphic platforms).脉冲神经网络(spiking neural network, SNN)作为神经形态平台的基本计算范式,使用离散的脉冲来传递信息,这使得它们在信息处理和能耗效率方面更接近于生物神经系统,并在模式识别、感知系统等应用中表现出色[4-6].为模拟大脑的操作动力学,神经形态平台通常包含数千个并行核心.例如,IBM 在其深蓝超级计算机平台上使用了近 150 000 个 CPU 来模拟猫的大脑皮层模型[7].SpiNNaker 项目[8]已经实现了一个具有惊人的 1 036 900 个 ARM 内核的大规模神经形态系统,这就导致网络中充斥着海量的数据包.因此,在神经形态平台上运行 SNN 的一个关键挑战是如何在数千个核心之间分配大量的数据包流量.传统的芯片通信架构使用总线结构,其中所有模块共享一个通信总线.随着芯片集成度的增加和系统规模的扩大,总线架构面临着通信带宽瓶颈、延迟增加和功耗上升等问题.片上网络(network-on-chip, NoC)应运而生,为这些问题提供了一种有效的解决方案.NoC 是一种在芯片级别上实现高度集成的通信架构,可以提供更高的吞吐量、更低的延迟以及良好的可扩展性、可重构性、灵活性.这些特性使得 NoC 成为目前神经形态平台常采用的一种互连体系结构.

目前神经形态平台 NoC 的设计还面临着一些挑战.首先在传输模式方面,一个神经元的脉冲信号常常需要传递给多个有突触连接的目标神经元.要处理这种一对多的传输模式,多播技术比传统的单播技术更加适合[9].多播技术可以一次性将脉冲信号发送给所有目标神经元,而不需要逐个发送,这显著提高了数据传输的效率,减少了通信开销和延迟.此外,多播更逼真地再现了生物神经网络的并行处理能力和信息传播方式[10].然而之前的很多神经形态平台,如 TrueNorth[11]、Loihi[12]、天机芯[13]等,只支持单播路由策略.这种通过多次点对点传输来完成多播的方式会导致额外的功耗开销和低传输效率.除了路由器的传输模式以外,NoC 及其路由策略的可扩展性和存储信息开销也是满足未来神经形态计算需求的关键之处.部分工作[14-15]基于目的驱动的路由方式,将所有的目的地路由信息以分组报头的形式嵌入到数据包中.这种方式会严重影响 NoC 的可扩展性,当目的地数量增多时,数据包将会呈线性膨胀;还有一些工作[16-18]基于源驱动的路由方式,将所有可能的路由路径提前计算好,并保存在本地存储器中,数据包中只传输其源地址.这种方式存在的一个问题是经过每个中间节点都需要检索路由表,延迟和功耗代价过大.

为了解决这些问题,本文针对大规模神经形态平台的通信需求,提出了一个脑启发的 NoC 设计.本文的主要贡献包括 3 个方面:

1)借鉴大脑的小世界网络特性,提出了自适应、无死锁的区域广播(region-broadcast, ReB)路由方案,通过统一的数据包格式和路由策略将单播、多播和广播融合起来,以更加通用简洁的方式适应神经形态平台的各种数据通信模式,并且设计了支持多种脉冲传输模式的路由器.ReB 利用脉冲发放的空间局部性,获得了更加均衡的链路负载和更低的峰值流量.

2)设计了一个突触连接索引方法用来替换传统的路由表,将神经元之间的突触连接信息存储起来,从而使 NoC 获得良好的可扩展性和较低的传输功耗.

3)在合成流量、4 个 SNN 应用以及脑皮质柱网络上对 ReB 和其他相关工作进行了链路负载、延迟、吞吐量和功耗等方面的对比,展示了在神经形态平

台 NoC 的设计中，我们的 ReB 路由策略和突触连接索引方法会更加具有优势.

实验表明，与现有工作相比，ReB 路由方案将峰值脉冲流量和链路负载标准差分别降低了 11.5% 和 20.4%，并且有效提升了网络的延迟、吞吐量和功耗等方面的性能. 同时，所提出的 ReB 路由器的带宽达到 0.24 spike/cycle，硬件实现面积仅为 0.014 mm^2.

1 背景及相关工作

本节将介绍大脑网络和神经元之间通信相关的背景知识，以及概述一些已有的神经形态平台 NoC 和专用 NoC 设计.

1.1 大脑的小世界网络特性

随着从人类神经影像数据中重构网络技术的发展，Latora 等人[19] 的工作表明人类大脑网络有小世界特性，具有高度聚集的结构和全局性的高效率，即神经元倾向于形成紧密联系的群体，而且具有较短的特征路径长度. 图 1 展示了大脑区域之间的连接模式.

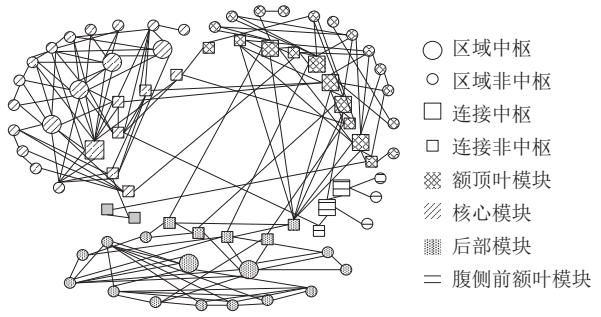


Fig. 1 Modular structure of brain network[2]

图 1 大脑网络的模块化结构[2]

由于连接远距离神经元的布线成本昂贵，大脑网络的进化倾向于将连接的神经元尽可能靠近，这有助于提高大脑网络的拓扑效率和鲁棒性[2].

真实生物神经元的平均放电速率约为 10 Hz. 在人类大脑中，10^{11} 个神经元的集体活动导致每秒产生多达 10^{12} 数量的脉冲信号. 尽管中枢神经系统内的神经元之间存在大量的尖峰信号，但大脑仍然通过其模块化结构实现了显著的能量效率. 这种非凡的效率赋予大脑在认知和感知等复杂任务中强大的学习和记忆能力. 大脑网络通信成本有：延迟、能量和信息成本 3 个方面[20]. NoC 的通信模型在这 3 个方面之间实现良好的平衡，有可能构成生物学意义上更加合理的神经通信策略.

1.2 神经形态平台 NoC 及专用 NoC

在大规模神经形态平台研究领域，已经涌现了大量的 NoC 设计. 2014 年 IBM 发布的大规模仿人脑芯片 TrueNorth[11] 含 4 096 个核心，基本计算核通过 2D mesh 路由形成可拓展的大规模神经形态网络. 2019 年清华大学的异构融合类脑芯片天机芯[13] 具有高速度、高性能、低功耗的特点，“天机芯 I”采用了 2D mesh 路由结构，每块芯片含有 6 个 FCore，片内的 6 个 FCore 呈 2×3 矩形 2D 排列，每个 FCore 的路由节点可以与东南西北 4 个方向的 FCore 交换信息，同时也可以将信息发送至本地 FCore. Intel 的多核类脑芯片 Loihi [12] 实现了异步片上网络，用于核间通信，网络拓扑结构为 2D mesh，由于与芯片间消息事务相关的死锁保护的原因，Loihi 的 NoC 使用 2 个独立的物理路由器网络. 斯坦福大学研制的 Neurogrid[21] 数模混合系统中每个 256×256 的神经元阵列组成一个神经核，16 个神经核通过树形的拓扑连接结构形成分级网络. Neurogrid 采用多播树的片上路由策略，通过分 2 个阶段（向上阶段和向下阶段）进行路由，并将数据包的分流限制到向下阶段，从而避免死锁. 但这种路由方法与其独特的神经核树形拓扑连接方式相适应，不具备通用性. SpiNNaker[22] 是一种专门设计用于支持大脑中各种通信的计算机，它的关键创新是通信架构，路由系统采用基于 2D triangular mesh 的轻量级多播包路由机制，并提出了增强的最短路径路由（enhanced shortest path routing, ESPR）和邻居探索路由（neighbour exploring routing, NER）这 2 种源驱动的路由方式来完成 SpiNNaker 的通信需求. 2021 年复旦大学设计的负载感知多播路由（load-aware multicast routing, LAMR）策略[16] 主要采用源驱动的基于树的多播路由机制. LAMR 通过宽度优先搜索来寻找拥塞较小的路由路线，并且尽可能合并多播路径，实现了较高的通信效率.

此外，近年来还出现了许多专用的 NoC 设计. 2017 年 Akbari 等人[23] 设计了一种用于神经形态结构的高性能片上网络拓扑——Dragonfly，以解决传统拓扑结构具有高直径、低平分带宽和较差的集体通信支持. 2023 年 Yazdanpanah[14] 提出了基于 2 层 2D mesh 拓扑结构的 TLAM 路由方法，采取了混合树/路径的多播路由策略. 它将片上网络划分成 4 个分区，每个分区有 1 个中心节点，并且通过 4 条新增链路将中心节点连接起来. 单播数据包采用基于转向的无死锁 XY 路由，多播数据包在每个本地分区采用哈密顿路径进行路由. 杨智杰等人[24] 设计的类脑处理器异步片上网络架构 NosralC 采用异步链路和同步路由器实现，在资源和性能开销不大的前提下，实现了延迟

的降低和能效的提升.

目前大多数神经形态平台 NoC 只支持单播路由策略,通过多次单播来实现多播. 专用 NoC 设计虽然在一定程度上解决了传输模式单一的问题,但路由策略仍然存在不够灵活高效的问题,在支持多播方面,对于网络的死锁避免问题也没有很好的解决办法. 此外,路由信息存储方法仍然存在可扩展性低和功耗开销过大的问题.

2 脑启发的 NoC 设计

在本节中,我们主要介绍针对大规模神经形态平台片上互连通信所设计的 NoC 架构,包括整体的设计方案、详细的路由策略、路由器架构和突触连接索引方法,并对 NoC 的可扩展性进行了分析讨论.

2.1 整体设计方案及路由策略

受人脑小世界网络特性的启发,本文提出了一种用于大规模神经形态平台上脉冲数据包传输的无死锁 ReB 路由策略. ReB 路由策略可以支持单播、多播和广播的并行路由计算和传输. 路由算法的整体思路如图 2 所示. 详细的流程解释如下:

1)根据 SNN 模型中神经元的互连情况,将连接紧密的神经元尽可能划分到 1 个计算核心,完成神经元到 NoC 的映射.

2)针对每个源神经元的目标神经元,我们通过算法(如层次聚类算法)对目标神经元进行聚类,得到一些簇集. 对每个聚类簇用矩形框起来,并用一条对角线的 2 个坐标点表示: $A(x_1, y_1)$, $B(x_2, y_2)$, 其中 A 是矩形的左上角顶点, B 是右下角顶点.

3)将每个聚类簇矩形信息,即 A 和 B 这 2 个坐标点存储到与源神经元相对应的存储单元中. 当神经元产生脉冲事件时,它会访问本地存储器得到与该神经元相对应的聚类簇矩形,并将数据包发送到目标矩形区域的边界节点. 当到达目标区域后,数据包在该区域进行广播,区域内的节点访问突触连接信息以确定该数据包是否被接收.

在单播模式情况下,矩形区域是 1 个点;在多播模式情况下,多个目的神经元支持多个目的地和矩形区域;在广播模式情况下,目的神经元分布在 NoC 的所有核心上时,矩形区域包含整个 2D mesh 的 NoC. 图 2 的“区域广播路由”图展示了 ReB 路由策略的大致过程,它具有自适应和无死锁的特性. 通过 ReB 路由第 1 次到达矩形区域的节点,我们称之为过渡节点. 首先,ReB 路由在区域内进行广播的方式为:对于过渡节点,它会向所有方向(除去传入和边界方向)发送数据包;对于区域内的其余节点,从 X 方向进入的数据包在所有方向上(除去传入和边界方向)进行传输;从 Y 方向进入的数据包只沿着相同的 Y 方向发送. 其次,从源节点发往目标区域的过程是基于西向优先转向模型[25]进行路由,我们将其扩展以支持目的地不是单个节点而是一个矩形区域的情况:当源节点在矩形区域左边界的东边并且在区域的上方或下方时,数据包需要先向西传输到矩形区域的左边界的延长线上,再以最短距离传输到矩形区域. 使用这种方法的原因在于避免从区域外的单播切换到区域内的广播时,发生违背转向规则的传输,正如图 3 所示.

在 NoC 中,数据包需要在各个处理单元之间不断传输. 死锁情况会导致数据包在某个节点或路径上无法前进,从而阻塞整个网络的通信. 随着片上核心数量的增加,通信路径变得更加复杂. 如果网络中存在死锁风险,系统的扩展会变得更加困难. ReB 路由策略通过禁止由南向西和由北向西 2 个方向的转向,使得在 NoC 上不会形成环路依赖. 即使对于多播数据包的传输,也可以避免在 NoC 上发生死锁的现象.

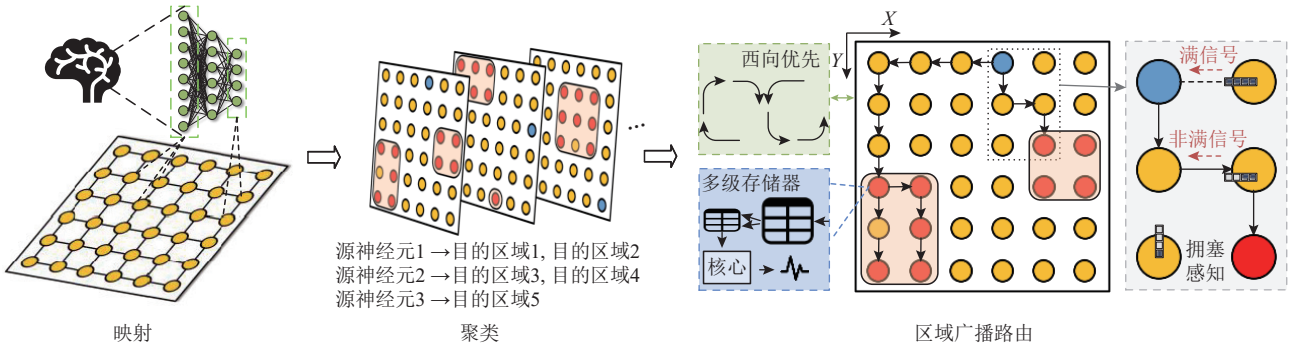


Fig. 2 The overall flow chart of ReB routing strategy

图 2 ReB 路由策略的整体流程图

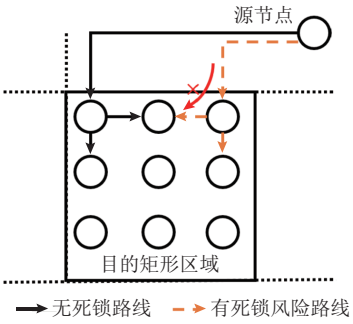


Fig. 3 An illustration of deadlock avoidance of ReB

图 3 ReB 的死锁避免示意图

本文提出的路由机制还实现了拥塞感知的路径选择. 如图 2 右图所示，路由器在收到东向链路上的满信号时，会将数据包向南发出，而不是阻塞在当前节点，这种自适应的路径选择发生在图 4 中的位置 1 和位置 2.

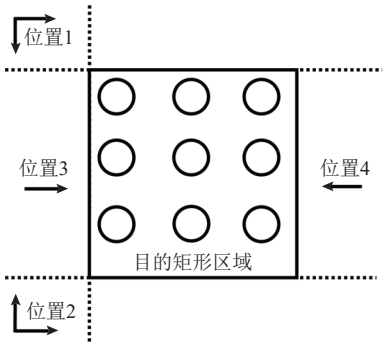


Fig. 4 Routing directions allowed by ReB strategy in different situations

图 4 在不同情况下 ReB 策略允许的路由方向

对于源节点在位置 1 的情况，可以根据相邻节点输入信道的满信号，选择向东或者向南传输；源节点在位置 2 的情况同理，选择向东或向北传输. 对于源节点在位置 3 和位置 4 的情况，数据包直接沿 X 方向传输到目的矩形区域即可. 由于这些情况都是以最短路径传输到目的区域的过渡节点上，因此也不存在活锁的发生. 具体路由算法的伪代码见算法 1.

算法 1. ReB 路由算法.

输入：当前节点坐标(x, y)，目标区域的左上顶点坐标(X_L, Y_L)，右下顶点坐标(X_R, Y_R)，数据包传入方向 *inport*；

输出：数据包的所有输出方向 *outs*.

① if(x, y)在目标区域内/* 在目标矩形区域内*/

② if 数据包是第 1 次到达目标矩形区域或者 *inport*==east or west

③ *outs*←除去 *inport* 和边界方向的其他所有方向；

④ else if *inport*==north or south

⑤ *outs*←*inport* 的反方向；

⑥ endif

⑦ else /* 在目标矩形区域外*/

⑧ if $x > X_L$

⑨ *outs*←west;

⑩ else

⑪ if $Y_L \leqslant y \leqslant Y_R$ 或者东边没有传来满信号

⑫ *outs*←east;

⑬ else if $y < Y_L$

⑭ *outs*←south;

⑮ else if $y > Y_R$

⑯ *outs*←north;

⑰ endif

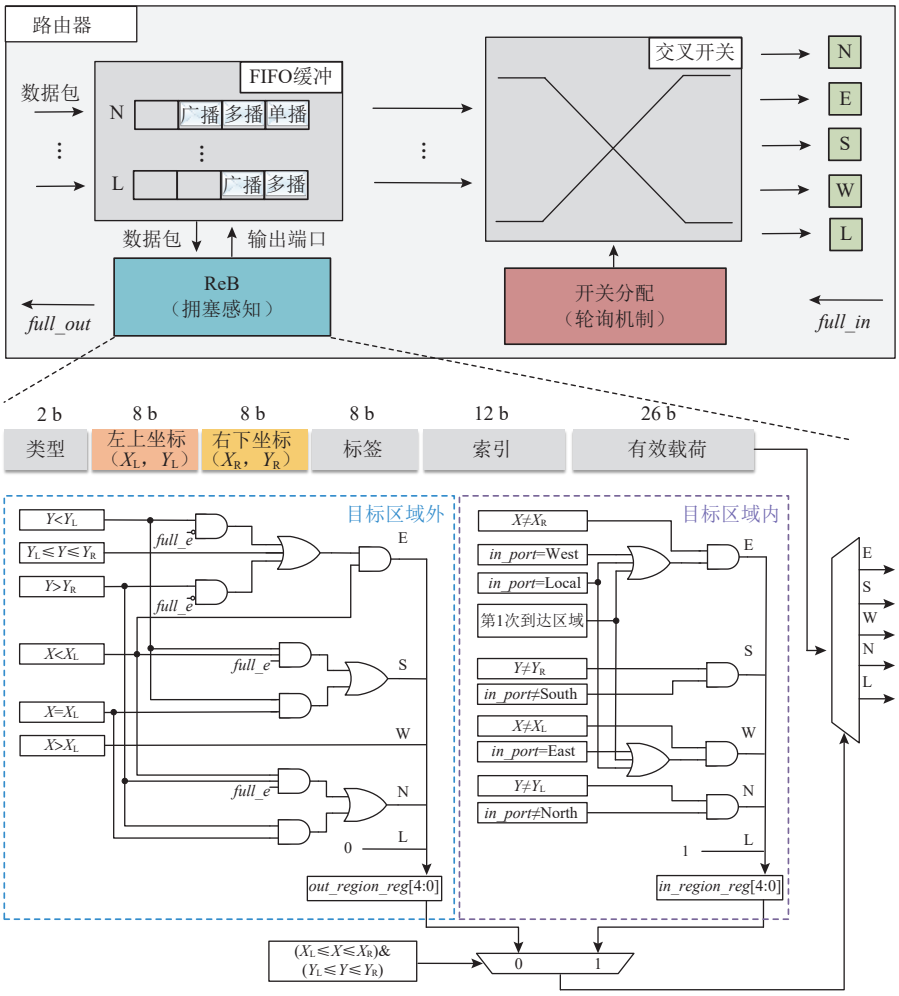
⑱ endif

⑲ endif

2.2 路由器的硬件设计

本文还实现了支持 ReB 策略的路由器，用于混合模式的脉冲数据包传输. 图 5 展示了路由器的硬件整体架构. 为了进一步减少面积开销，我们不使用虚通道技术，也不为输出端口配备缓冲区. 当数据包到达输入端口时，它会被写入相应的先进先出（first in first out, FIFO）存储器. 然后，ReB 逻辑执行路由计算，以确定此数据包的输出端口. 路由输出端口根据来自相邻路由器输入信道的满信号自适应地改变，这可以有效地减少 NoC 中的拥塞. 接下来，数据包将进入开关分配阶段，它会仲裁开关的输入输出端口. 一旦分配到输出端口，单播数据包就会从缓冲区中读出并进行到开关传输阶段. 最后，数据包经过链路传输到下游路由器. 对于多播或者广播数据包，它可能会请求多个输出端口，数据包将被复制到多个输出端口. 只有当数据包被成功传输到所有的目的端口后，它才会从 FIFO 中被移除.

数据包大小统一为 64 b，最高位的 2 b 表示数据包类型，包括单播（00）、多播（01）、广播（10）. 接下来的 2 个 8 b 字段分别表示目的矩形区域的左上坐标(X_L, Y_L)和右下坐标(X_R, Y_R). 标签 *tag* 和索引 *index* 字段用于判断矩形区域中的节点是否接收该数据包，在目标区域之外不发挥作用. 最后一个字段表示数据有效载荷 *data*. 图 5 还展示了所提出 ReB 路由策略的硬件电路设计. 该电路由基本门电路、比较器和多路选择器组成. 比较器用于将当前路由器地址与目标地址进行比较，比较的结果和邻居路由器传来的满信号综合起来作为输出方向的判定条件. 多路选



E: East; S: South; W: West; N: North; L: Local

Fig. 5 The microarchitecture of the router

图 5 路由器的微架构

择器根据当前路由器位于矩形区域内外的情况对输出方向进行一个选择.

2.3 突触连接索引方法

在本文中，我们部署了一种突触连接索引方法来配合 ReB 路由策略并替换掉多播路由表，使得 NoC 的信息传输可以覆盖每个核心的所有神经元，相当于模拟真实神经元之间的突触连接. 当神经元发放脉冲时，神经元 ID 将用于查找目的地信息. 具体流程如图 6 所示. 源神经元 ID 被用作访问发送端第 1 级存储器的索引，以获得多播区域的基地址和区域数量. 然后根据基地址和区域数量访问发送端第 2 级存储器的相应条目. 每个条目都包括目标区域的左上坐标、右下坐标、标签和索引字段. 条目中字段的含义与数据包格式中相应字段的含义一致. 最后，这些与目的神经元相关的基本信息被组装成数据包并由路由器发送到网络中. 当脉冲数据包到达目的节点处的路由器时，数据包的索引字段用于检索接收端第 1 级存储器，将脉冲数据包中的标签字段与索引条目中的标签进行比较，如果它们匹配，则认为该数据包有资格由当前核心接收. 接着，路由器使用基地址和索引条目中的神经元数量来检索接收端第 2 级存储器中所有的目标神经元 ID 和树突 ID. 这些 ID 用于访问核心数据区中的突触权重，为神经元计算做准备. 发送端和接收端的多级存储器构建了不同神经元之间突触连接的桥梁. 图 6 还展示了一个在 NoC 上的不同核心之间神经元互连通信的示例，核心 0 中的 1 个神经元分别与核心 1 中的神经元 1、核心 2 中的神经元 2、核心 3 中的神经元 3 和神经元 4 存在突触连接. 当源神经元发送脉冲数据包时，它通过发送端的 2 级存储器获取目的区域的路由信息，按照 ReB 路由策略传输到核心 1, 2, 3 所组成的目的区域. 当核心 3 收到该脉冲数据包时，数据包会访问接收端的多级存储器，通过索引和标签字段判断出当前核心可以接收该数据包，进而获取要传输到的

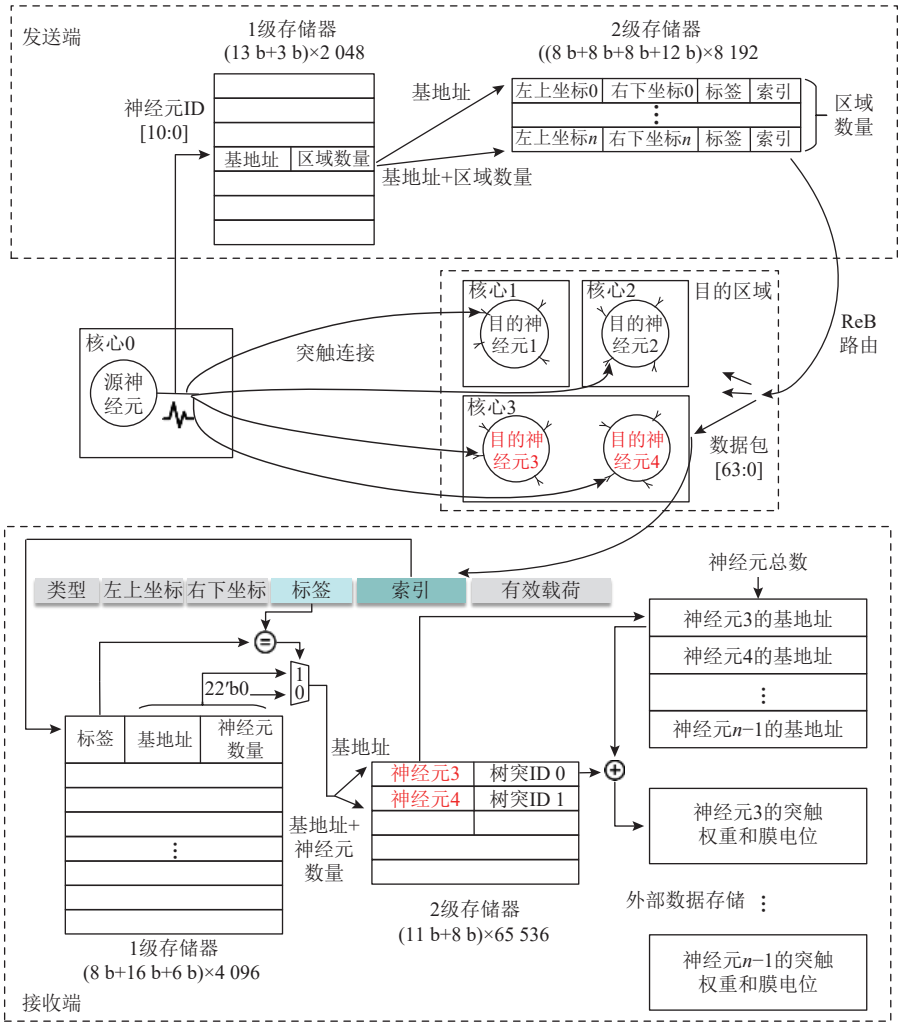


Fig. 6 Illustration of the synaptic connections indexing method

图 6 突触连接索引方法的示意图

具体神经元 ID, 即神经元 3 和神经元 4.

2.4 可扩展性分析

所提出 NoC 具备良好的可扩展性. 一方面, 数据包采用统一的格式来完成混合模式的脉冲数据传输. 多播目的地以矩形区域的形式嵌入到大小为 $O(\log N)$ 的 single-flit 数据包中, 其中 N 表示 NoC 的核心数. 由于数据包无需存储所有的目的地路由信息, 这大大节省了信息传输开销. 矩形区域的编码长度随着片上核心数目的增加呈对数变化, 这对于大规模神经形态平台的扩展具有很大的优势. 此外, mesh 拓扑结构本身就具备灵活扩展的特性, 这种扩展不会显著增加整体复杂度或影响网络性能.

另一方面, 目的驱动的路由方式在传输路径上不需要路由表. ReB 路由只需要在源节点和目的区域索引突触连接关系来发送和接收数据包, 而不是路径上的每一跳都要访问路由表. ReB 路由访存功耗与源驱动的路由方式所产生的访存功耗的比值为

$$\frac{P_{\text{ReB}}}{P_{\text{源驱动}}}=\frac{(1+k)\times e}{(1+h)\times e}=\frac{1+k}{1+h}, \tag{1}$$

其中 e 表示 1 次访存所产生的能耗, k 表示多播的目的地数量, h 表示传输的总跳数. 如图 7 所示, 以基于源驱动的维序路由[18]为例, 我们统计了不同 NoC 拓扑大小下源驱动路由的平均总跳数, 并计算了式(1)

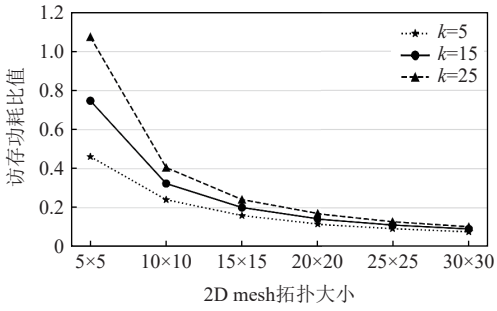


Fig. 7 Memory access power consumption ratio under different sizes of NoC

图 7 不同 NoC 规模下的访存功耗比值

中访存功耗的比值. 当 2D mesh 的拓扑变大时, 对于随机产生目的地的方式来说, 该比值会随之变小. 主要原因是 ReB 路由的访存功耗与拓扑规模大小无关, 仅与多播目的地数量相关, 而源驱动路由中的总跳数 h 与网络拓扑大小呈正相关.

ReB 路由策略的良好可扩展性对于大规模神经形态平台至关重要, 它能够确保系统在面对不断增加的计算需求时, 可以灵活地增加计算核心或存储资源, 并且仍然保持高效的性能和较低的功耗, 以应对更大规模的复杂神经网络任务.

3 实 验

本节分析讨论了所提出的 ReB 路由策略对链路负载、网络延迟和吞吐量的影响, 以及所提出的突触连接索引方法对功耗的影响. 这些都是衡量 NoC 性能的关键指标. 本节将基于 PAT-Noxim[26] 仿真器展开实验. 它是基于 SystemC 开发的周期精确的 NoC 模拟器, 可以进行延迟、吞吐量等数据的分析, 也可以根据有关功耗的细粒度统计数据进行性能的评估. 此外, 我们选取了一些经典的路由策略以及最近一些 NoC 工作中所采取的路由策略作为对比, 包括 UNICAST, XY[25], ESPR[18], LAMR[16]. 其中 UNICAST 算法通过多次的单播来完成多播, 作为对比的基线策略, 以衡量多播技术所带来的性能提升; XY 多播路由优先往 X 方向传输数据包, 遇到 Y 方向上存在的多播目的地时复制数据包并产生路由分支; ESPR 是 SpiNNaker 神经形态平台所提出的片上多播路由策略, 它通过尽可能合并多播的路由路径来减少链路负载; LAMR 是复旦大学团队设计的基于树的多播路由机制, 它通过宽度优先搜索来寻找拥塞较小的路由路线. 这些策略都在仿真器上进行了部署和实现. 为了使实验结果更具可靠性, 我们进行重复实验计算性能指标的平均值. 具体的仿真实验参数配置如表 1 所示.

此外, 为了更好地发挥所提出 ReB 策略的自适应路由的优势和减少由于西向优先产生的额外跳数, 我们对 ReB 的多播随机目的地映射进行调整, 将源节点尽量放置在图 4 中相对目的区域的位置 1~4, 这种方式命名为 ReB-MA (mapping adjustment).

3.1 合成流量模式下的性能评估

为了评估 ReB 和其他路由策略的链路负载、延迟和吞吐量等性能指标, 我们使用了多种合成通信流量模式, 包括随机(random)、翻转(transpose)、热点(hotspot).

Table 1 Simulation Parameters Setting

表 1 仿真实验参数配置

参数	取值
mesh 大小	10×10
FIFO 深度	8
微片大小/b	64
数据包长度	signle-flit
交换机制	虫洞
路由器流水线长度	4 级
合成流量模式	随机, 翻转, 热点
多播目的地数量	10, 20, 30
注入率/(packet/(cycle·node))	0.01~0.055
热身时间/cycle	1 000
运行时间/cycle	20 000

3.1.1 链路负载

图 8(a)为各种路由方案下所有链路流量分布的盒图, 可以看出由于更好地利用了目的地位置的局部趋向性, ReB 和 ReB-MA 策略的流量分布更加集中均衡. 与 LAMR 策略相比, ReB-MA 的峰值流量降低了 11.5%. UNICAST 通过多次单播来完成 1 次多播的数据包传输, 这种方式显著地增加了网络流量, 并且降低了传输效率. ESPR 和 LAMR 策略通过一些方法来合并不同数据包的多播路径, 从而减少了传输过程中的多播路由分支. 因此, 相比于没有路径合并机

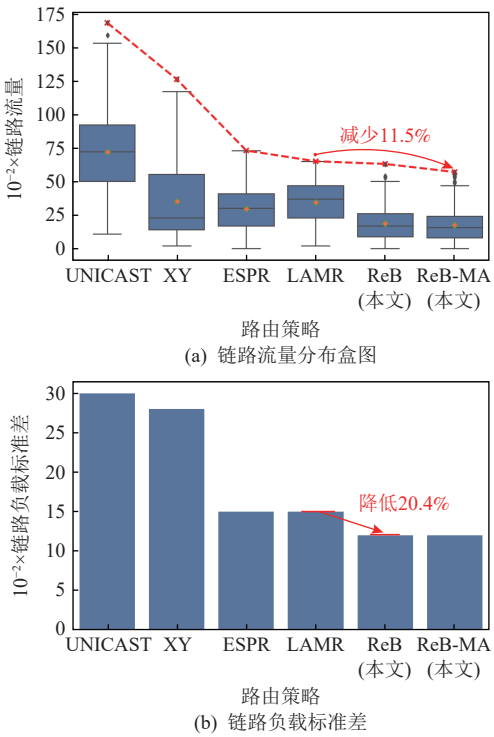


Fig. 8 Link traffic distribution and link load standard deviation

图 8 链路流量分布以及链路负载标准差

制的 XY 多播来说，平均链路流量和峰值负载都得到了很大程度的降低. 此外，由于映射调整后的 ReB-MA 可以更多地利用路由的自适应拥塞感知策略去改变数据流量的传输路径，网络流量相比于 ReB 也会更好地分散在不同的链路上，从而进一步降低链路的峰值负载.

如图 8(b)所示是所有链路负载标准差的对比实验结果. 标准差δ可以通过式(2)(3)得到.

$$\delta=\sqrt{\frac{\sum_{i=1}^{m}(x_i-\mu)^2}{m}},\tag{2}$$

$$\mu=\frac{\sum_{i=1}^{m}x_i}{m},\tag{3}$$

其中，i 表示链路编号，x_i 是链路 i 的数据流量，m 是有效链路的总数. 在这个例子中，对于 10×10 大小的 2D mesh NoC 来说，m=360. 我们从图 8(b)中可以发现，采用 ReB 路由的链路负载的标准差减少了 20.4%，获得了更加均衡的负载流量，可以在很大程度上缓解网络上的拥塞.

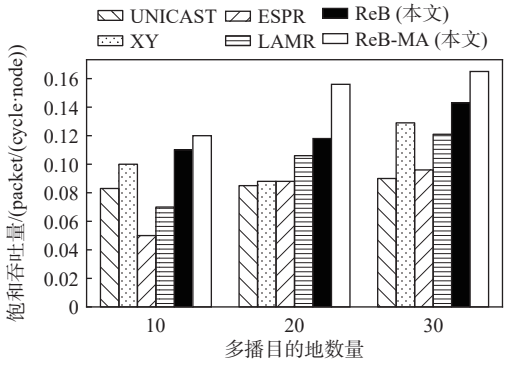
3.1.2 吞吐量

如图 9 所示，我们研究了在不同多播目的地数量和不同拓扑大小下的网络饱和吞吐量，以表征系统的整体数据传输能力. 饱和吞吐量的定义为：

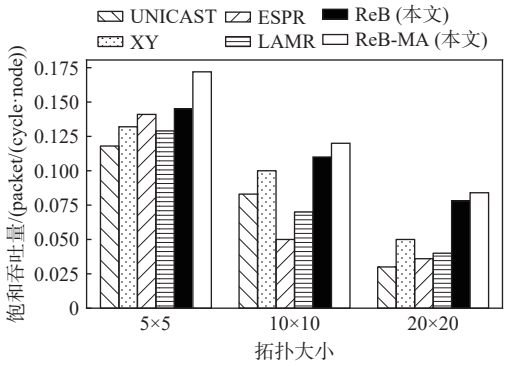
$$\text{吞吐量}=\frac{\text{接收的数据包总数}}{\text{仿真运行时间}\times\text{节点数}}.\tag{4}$$

我们逐渐增加注入率，直到网络吞吐量不再增加，此时的吞吐量被定义为饱和吞吐量. 从图 9(a)可以看到，从 10 个多播目的地到 30 个多播目的地，ReB 和 ReB-MA 策略的网络饱和吞吐量逐渐增加，相对于其他路由策略，始终保持较大的饱和吞吐量. 在 30 个多播目的地时，ReB-MA 策略下每个路由器的饱和吞吐量可以达到 0.16 packet/(cycle·node). 我们设计的无死锁机制使得在注入率较大时，网络依然可以连续不断地对数据包进行传输. ESPR 和 LAMR 路由策略由于没有相应的死锁避免机制，相比于无死锁的 XY 多播路由，并没有取得吞吐性能上的优势.

NoC 的拓扑大小取决于具体的神经形态规模需求，较大的 NoC 会带来传输跳数和延迟的增加，较小的 NoC 会导致系统支持的神经元数量减少. 图 9(b)展示了在不同拓扑规模下的网络饱和吞吐量，可以看到随着拓扑规模的变大，平均传输路径变长，UNICAST 单播方式的饱和吞吐量下降趋势尤为明显，其他 5 种多播算法也出现不同程度的下降. ReB 和



(a) 不同多播目的地数量下的饱和吞吐量



(b) 不同拓扑大小下的饱和吞吐量

Fig. 9 Saturation throughput for different destination numbers and topology size

图 9 不同多播目的地数量和拓扑大小下的饱和吞吐量

ReB-MA 策略在不同拓扑大小下保持相对较高的饱和吞吐量，在 20×20 的拓扑大小下，饱和吞吐量仍然达到 0.08 packet/(cycle·node).

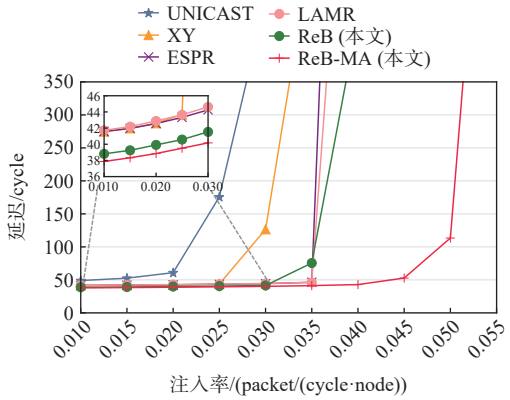
3.1.3 平均延迟

图 10 展示了不同数据包注入率下在随机、翻转、热点这 3 种不同的合成流量模式下不同的路由策略. 延迟被定义为数据包到达目的地被接收的时间戳与数据包产生的时间戳之间的差值.

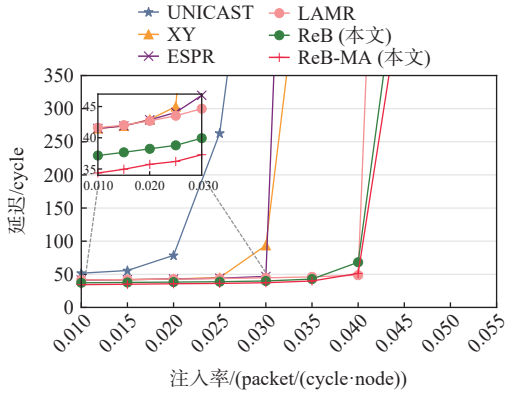
由图 10 可以看到，对于不同的流量模式，ReB 和 ReB-MA 在轻到中等流量负载下都表现出相对较低的延迟. 相较于完全单播的方式，ReB 路由在 0.01 注入率下，平均延迟减少了 20.7%~30.8%. 此外，在随机和热点的流量模式下，本文方案可以支持更高的数据流量注入率，并且在 NoC 上不会产生死锁和活锁的现象，这对于大规模神经形态平台中的信息传输具有重要意义. 虽然在翻转流量模式下，LAMR 路由表现较为出色，但其在高注入率下很容易发生死锁，导致系统运行被阻塞，需要重复运行很多次仿真来获取实验结果.

3.2 真实 SNN 应用下的性能评估

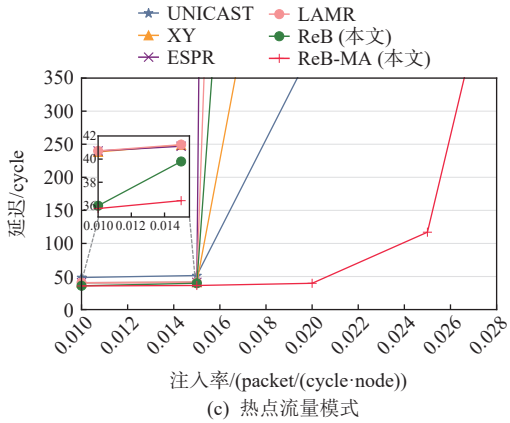
SNN 作为神经形态平台的基本计算范式，将其



(a) 随机流量模式



(b) 翻转流量模式



(c) 热点流量模式

Fig. 10 Average latency changes for different injection rates

图 10 不同注入率下的平均延迟变化

进行映射部署到 NoC 上进行仿真验证是有意义的. 本实验使用了 4 种不同类型的 SNN 应用, 它们具有不同的连接模式和数据集, 如表 2 所示.

QTDB 和 Ev-object 数据集对应的 SNN 网络结构都只有全连接, 因此多播比例是 100%. 除了全连接外, MNIST 和 DVS-Gesture 对应的 SNN 网络还有卷积层和池化层. 图 11 展示了在不同 SNN 应用下的归一化运行时间和功耗开销.

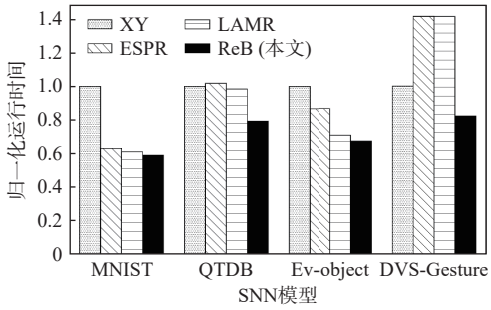
ReB 路由策略利用神经元高度聚集结构的特性, 在更短的运行时间内处理完所有输入的脉冲信号.

Table 2 SNN Applications

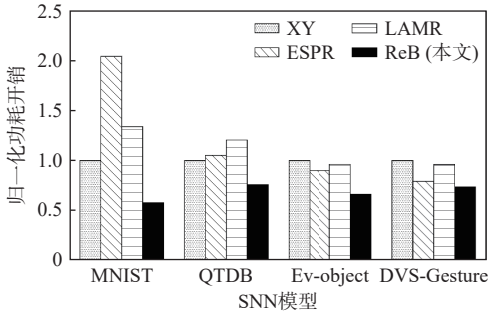
表 2 SNN 应用程序

数据集	网络结构	平均数据包数量	平均多播比例/%
MNIST[27]	4C3-MP2-4C3-MP2-128FC-10FC	14 632	37.90
QTDB[28]	36H-6FC	205 208	100.00
Ev-object[29]	128FC-256FC-36FC	95 700	100.00
DVS-Gesture[30]	MP2-16C3-16C3-MP2-16C3-MP2-{16C3}×3-64FC-16FC-5FC	1 797 399	60.91

注: 4C3 表示卷积通道数为 4、卷积核大小为 3×3, MP2 表示步长为 2 的最大池化层, 10FC 表示输出维度为 10 的全连接层, 36H 表示 SRNN 网络结构里的 36 个隐藏层神经元, {*K*}×3 表示有 3 个 *K* 结构.



(a) 不同SNN应用下的归一化的网络运行时间



(b) 不同SNN应用下的归一化的功耗开销

Fig. 11 Performance comparison with various SNN applications

图 11 不同 SNN 应用下的性能对比

突触连接索引方法节省了在路由路径上查找路由表所产生的功耗, 而只需要在源节点和目的区域内对突触连接的存储信息进行访问. 这对于网络中较长的传输路径来说, 具有很大的能效优势.

3.3 脑网络仿真

目前人类对大脑的认知程度还十分有限, 但已经涌现了大量的脑仿真建模工作. 为了更加全面地评估 ReB NoC 架构的性能, 我们基于 Potjans 等人[31]所提出的分层皮质微环路模型开展仿真验证实验, 如图 12 所示. 皮质柱被认为是大脑的基本功能块, 第 2/3, 4, 5, 6 层分别由模型神经元的兴奋性和抑制性群体表示, 群体的输入由丘脑皮层输入靶向层 4 和 6 以及所有群体的其他外部输入表示, 模型尺寸对应于 1mm^2 表面下的皮质网络.

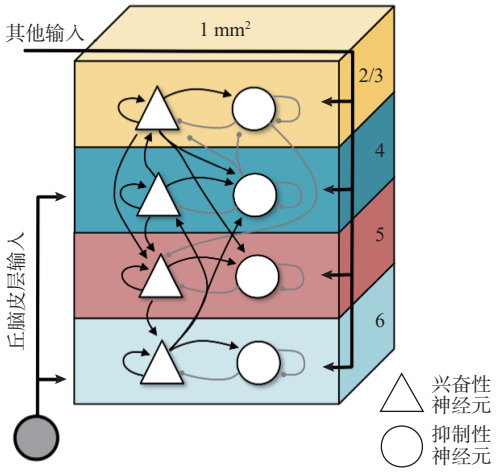


Fig. 12 Layered cortical network model[31]

图 12 分层皮质网络模型[31]

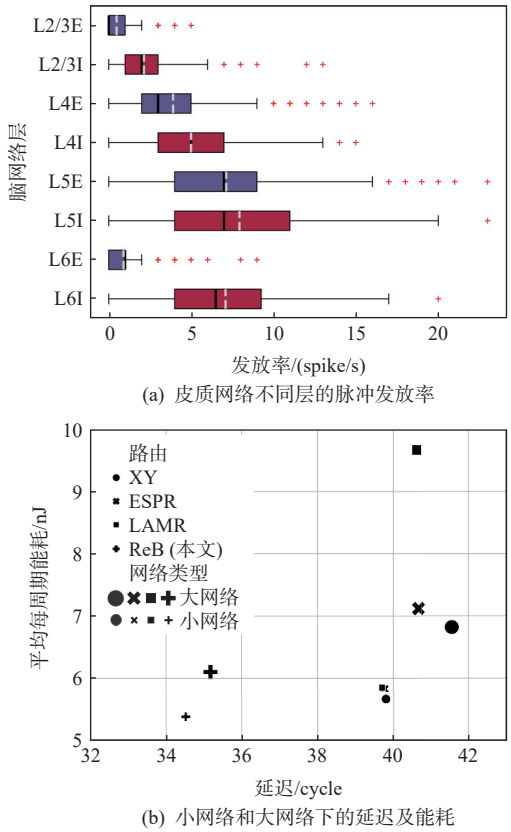


Fig. 13 Performance evaluation under layered cortical networks models

图 13 分层皮质网络模型下的性能评估

图 13(a)统计了该脑网络模型中每个神经元群的脉冲发放频率，L5 层的脉冲发放较为频繁，意味着会输出更多的数据包到网络中. 各个神经元群的平均脉冲发放率不超过 10 spike/s，每个计算核心可以容纳 2 048 个神经元，因此一个计算核心的最大发放率为 20 Kspike/s. 本文提出的 NoC 路由器带宽可达 0.24 spike/cycle，在 100 MHz 条件下每秒可以处理 24×10^6 的脉冲，因此满足脑网络仿真的吞吐需求，并且可以将脑网络的仿真时间缩短约原来的 1/1 000.

我们基于该模型生成了 5 015 个神经元的小网络和 38 586 个神经元的大网络用于仿真实验. 图 13(b)统计了 2 种脑网络规模下的延迟和能耗数据，并与其他 3 种多播 NoC 进行比较. 得益于 ReB 路由的脑启发多播特性和突触连接索引方法，ReB 路由策略在不同规模的分层皮质网络下都具有较低的传输延迟和能耗开销，可以有效支持脑网络应用下的脉冲信息传输.

3.4 与相关先进工作的比较

本文提出的路由器通过 Verilog HDL 实现，并基于 SMIC 28nm CMOS 工艺完成综合. 表 3 显示了 ReB 与其他先前工作中 NoC 的比较. 表 3 的上半部分列出了一些专用的 NoC 设计，它们通常用于一般片上系统(system on chip, SoC)中的数据传输需求，典型的应用场景包括处理器间的同步信号、缓存一致性协议等；表 3 的下半部分展示了一些用于神经形态平台的 NoC 设计，它们专注于满足神经网络计算的特定需求，通常包括 SNN 和 ANN 这 2 种应用场景，比如清华大学的异构融合类脑芯片 Tianjic[13] 在硬件上融合这 2 种神经网络的实现. 神经形态平台 NoC 上传输的数据类型包括脉冲数据和非脉冲数据，其设计核心在于高效的数据并行传输和多种数据类型的支持. 本文的 ReB 路由器的带宽可以达到 0.24 spike/cycle，硬件实现面积仅为 0.014 mm^2. 此外，只有 ReB

Table 3 Comparison of ReB and Related Advanced Work

表 3 ReB 与相关先进工作对比

神经形态平台	类型	工艺/nm	拓扑结构	路由算法	有无死锁避免	带宽/(spike/cycle)	面积/mm^2
MC-3DR[17]	专用 NoC 设计	45	3D mesh	混合模式，固定路由	无	0.250	0.031
TLAM[14]		45	分层结构	混合模式，固定路由	有	0.086	0.070
Tianjic[13]	神经形态平台的 NoC 设计	28	2D mesh	单播，固定路由	有		0.008
LAMR[16]		28	2D mesh	混合模式，自适应路由	无	0.016	
MerLin[32]		55	Fullerene-60	混合模式，自适应路由	无	0.200	0.013
ReB（本文）		28	2D mesh	混合模式，自适应路由	有	0.240	0.014

路由策略支持混合模式的脉冲数据传输，并且具备无死锁和拥塞感知的自适应路由机制.

4 结 论

本文提出了一种针对大规模神经形态平台设计的脑启发 NoC 方案. 所提出的无死锁 ReB 路由策略支持混合模式的传输和拥塞感知的路径选择. 突触连接索引方法为 NoC 架构提供了更好的可扩展性和更低的功耗. 实验结果表明，ReB 的 NoC 在延迟、吞吐量、信息成本和功耗之间实现了良好的平衡，为神经形态平台提供了一种更具生物学合理性的神经通信策略. 在未来，我们将基于该 NoC 研究更大规模的多层次众核互联架构和通信协议，能够容纳多种神经网络并支持任意神经元之间互联.

作者贡献声明：王智超负责方案设计、实验验证、论文撰写；陈亮负责理论指导、架构设计和论文修改；李千鹏负责论文修改和设计代码撰写；陈奥新负责实验数据记录及论文修改；刘昕负责代码设计；宋文娜负责实验验证.

参 考 文 献

[1] Lent R, Azevedo F A C, Andrade-Moraes C H, et al. How many neurons do you have? Some dogmas of quantitative neuroscience under revision[J]. European Journal of Neuroscience, 2012, 35(1): 1−9

[2] Bullmore E, Sporns O. The economy of brain network organization[J]. Nature Reviews Neuroscience, 2012, 13(5): 336−349

[3] Capra M, Bussolino B, Marchisio A, et al. Hardware and software optimizations for accelerating deep neural networks: Survey of current trends, challenges, and the road ahead[J]. IEEE Access, 2020, 8: 225134−225180

[4] Rathi N, Chakraborty I, Kosta A, et al. Exploring neuromorphic computing based on spiking neural networks: Algorithms to hardware[J]. ACM Computing Surveys, 2023, 55(12): 1−49

[5] Wu Jibin, Xu Chenglin, Han Xiao, et al. Progressive tandem learning for pattern recognition with deep spiking neural networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(11): 7824−7840

[6] Kumarasinghe K, Kasabov N, Taylor D. Brain-inspired spiking neural networks for decoding and understanding muscle activity and kinematics from electroencephalography signals during hand movements[J]. Scientific Reports, 2021, 11(1): 2486−2486

[7] Ananthanarayanan R, Esser S K, Simon H D, et al. The cat is out of the bag: Cortical simulations with 109 neurons, 1 013 synapses[C/OL]// Proc of the Conf on High Performance Computing Networking, Storage and Analysis. New York: ACM, 2009[2025-01-21]. https://dl.acm.org/doi/abs/10.1145/1654059.1654124

[8] Furber S B, Galluppi F, Temple S, et al. The SpiNNaker project[J]. Proceedings of the IEEE, 2014, 102(5): 652−665

[9] Vainbrand D, Ginosar R. Scalable network-on-chip architecture for configurable neural networks[J]. Microprocessors and Microsystems, 2011, 35(2): 152−166

[10] Kauth K, Stadtmann T, Brandhofer R, et al. Communication architecture enabling 100x accelerated simulation of biological neural networks[C/OL]//Proc of the Workshop on System-Level Interconnect: Problems and Pathfinding Workshop. New York: ACM, 2020 [2025-01-21]. https://dl.acm.org/doi/abs/10.1145/3414622.3431909

[11] Akopyan F, Sawada J, Cassidy A, et al. TrueNorth: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2015, 34((10):): 1537−1557

[12] Davies M, Srinivasa N, Lin T H, et al. Loihi: A neuromorphic manycore processor with on-chip learning[J]. IEEE Micro, 2018, 38(1): 82−99

[13] Pei Jing, Deng Lei, Song Sen, et al. Towards artificial general intelligence with hybrid Tianjic chip architecture[J]. Nature, 2019, 572(7767): 106−111

[14] Yazdanpanah F. A two-level network-on-chip architecture with multicast support[J]. Journal of Parallel and Distributed Computing, 2023, 172: 114−130

[15] Pu J, Goh W L, Nambiar V P, et al. A low power and low area router with congestion-aware routing algorithm for spiking neural network hardware implementations[J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2020, 68(1): 471−475

[16] Ding Chen, Huan Yuxiang, Jia Hao, et al. A hybrid-mode on-chip router for the large-scale FPGA-based neuromorphic platform[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2022, 69(5): 1990−2001

[17] Vu T H, Abdallah A B. Low-latency K-means based multicast routing algorithm and architecture for three dimensional spiking neuromorphic chips[C/OL]//Proc of the 2019 IEEE Int Conf on Big Data and Smart Computing (BigComp). Piscataway, NJ: IEEE, 2019[2025-01-21]. https://ieeexplore.ieee.org/abstract/document/8679363/

[18] Navaridas J, Luján M, Plana L A, et al. SpiNNaker: Enhanced multicast routing[J]. Parallel Computing, 2015, 45: 49−66

[19] Latora V, Marchiori M. Efficient behavior of small-world networks[J]. Physical Review Letters, 2001, 87(19): 198701

[20] Seguin C, Sporns O, Zalesky A. Brain network communication: Concepts, models and applications[J]. Nature Reviews Neuroscience, 2023, 24(9): 557−574

[21] Benjamin B V, Gao P, McQuinn E, et al. Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations[J]. Proceedings of the IEEE, 2014, 102(5): 699−716

[22] Furber S B, Lester D R, Plana L A, et al. Overview of the SpiNNaker system architecture[J]. IEEE Transactions on Computers, 2012, 62(12): 2454−2467

[23] Akbari N, Modarressi M. A high-performance network-on-chip topology for neuromorphic architectures[C]//Proc of the 2017 IEEE Int Conf on Computational Science and Engineering (CSE) and IEEE

Int Conf on Embedded and Ubiquitous Computing (EUC). Piscataway, NJ: IEEE, 2017: 9−16

[24] Yang Zhijie, Wang Lei, Shi Wei, et al. Asynchronous network-on-chip architecture for neuromorphic processor[J]. Journal of Computer Research and Development, 2023, 60(1): 17−29 (in Chinese)
(杨智杰, 王蕾, 石伟, 等. 类脑处理器异步片上网络架构[J]. 计算机研究与发展, 2023, 60(1): 17−29)

[25] Glass C J, Ni L M. The turn model for adaptive routing[J]. ACM SIGARCH Computer Architecture News, 1992, 20(2): 278−287

[26] Norollah A, Derafshi D, Beitollahi H, et al. PAT-Noxim: A precise power & thermal cycle-accurate NoC SImulator[C]//Proc of the 31st IEEE Int System-on-Chip Conf (SOCC). Piscataway, NJ: IEEE, 2018: 163−168

[27] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278−2324

[28] Laguna P, Mark R G, Goldberg A, et al. A database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG[C]//Proc of the Computers in Cardiology 1997. Piscataway, NJ: IEEE, 1997: 673−676

[29] Gu Fuqiang, Sng W, Taunyazov T, et al. Tactilesgnet: A spiking graph neural network for event-based tactile object recognition[C]//Proc of the 2020 IEEE/RSJ Int Conf on Intelligent Robots and Systems (IROS). Piscataway, NJ: IEEE, 2020: 9876−9882

[30] Zhang Yuan, Cao Jian, Chen Jue, et al. Razor SNN: Efficient spiking neural network with temporal embeddings[C]//Proc of the Int Conf on Artificial Neural Networks. Berlin: Springer, 2023: 411−422

[31] Potjans T C, Diesmann M. The cell-type specific cortical microcircuit: Relating structure and activity in a full-scale spiking network model[J]. Cereb Cortex, 2014, 24(3): 785−806

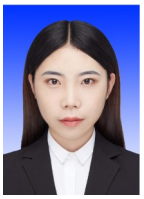
[32] Zhou Pujun, Yu Qi, Chen Min, et al. Fullerene-inspired efficient neuromorphic network-on-chip scheme[J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2023, 71(3): 1376−1380

Wang Zhichao, born in 2000. Master. His main research interests include neuromorphic computing and network-on-chip design.
王智超, 2000 年生. 硕士. 主要研究方向为类脑计算、片上网络设计.

Chen Liang, born in 1977. Master, associate professor, master supervisor. Member of CCF. His main research interests include design of computer architecture and integrated circuit, and brain-inspired computing.
陈　亮, 1977 年生. 硕士, 副研究员, 硕士生导师. CCF 会员. 主要研究方向为计算机体系架构与集成电路设计、类脑计算.

Li Qianpeng, born in 1999. Master. His main research interests include brain-inspired computing and brain-inspired processor. (liqianpeng2021@ia.ac.cn).
李千鹏, 1999 年生. 硕士. 主要研究方向为类脑计算、类脑处理器.

Chen Aoxin, born in 2000. Master. His main research interests include artificial intelligence and algorithm deployment optimization. (chenaoxin2022@ia.ac.cn)
陈奥新, 2000 年生. 硕士. 主要研究方向为人工智能、算法部署优化.



Liu Xin, born in 1995. Master, engineer. Her main research interests include integrated circuit design and computer architecture. (liuxin@ia.ac.cn)
刘　昕, 1995 年生. 硕士, 工程师. 主要研究方向为集成电路设计、计算机架构.

Song Wenna, born in 1989. Master, engineer. Her main research interests include system architecture and digital circuit design. (wenna.song@ia.ac.cn)
宋文娜, 1989 年生. 硕士, 工程师. 主要研究方向为系统结构、数字电路设计.