

基于人工神经网络的 NoC 智能动态链路管理方法

许威¹, 张霞²

(1. 北京城建设计发展集团股份有限公司, 北京 100037;

2. 北京理工大学信息与电子学院, 北京 100081)

摘要: 功耗是片上网络 (NoC) 主要限制因素, 链路状态的选择性开/关切换算法可降低电路级和系统级的链路功耗, 这些算法大多集中于一个简单的静态阈值触发机制, 该机制决定了是否应该打开或关闭链路; 为解决上述触发机制存在诸多限制, 提出了一种针对 NoC 的人工神经网络 (ANN, artificial neural network) 作为动态链路功耗管理方法, 该方法基于对系统状态的有监督在线学习, 通过使用小型可扩展的神经网络来关闭和打开链路, 从而提高预测能力; 基于人工神经网络的模型利用了非常低的硬件资源, 并且可以集成在大型网状和环面 NoC 中; 通过对不同网络拓扑上各种综合流量模型的仿真结果表明, 与静态阈值计算相比, 该方法在较低的硬件支出下可以节省功耗; 为解决链路管理 NoC 中的功耗问题提供思路。

关键词: 片上网络; 静态阈值; ANN 机制; 链路管理

NoC Intelligent Dynamic Link Management Strategy Based on Artificial Neural Network

XU Wei¹, ZHANG Xia²

(1. Beijing Urban Construction Design and Development Group Co., Ltd., Beijing 100037, China;

2. Beijing Institute of Technology School of Information and Electronics, Beijing 100081, China)

Abstract: Power consumption is the main limiting factor of network on chip (NOC). The selective on/off switching algorithm of link state can reduce the link power consumption at circuit level and system level. Most of these algorithms focus on a simple static threshold trigger mechanism, which determines whether the link should be turned on or off. In order to solve many limitations for the above trigger mechanism, an artificial neural network (ANN) for NOC is proposed as a method for the dynamic link power consumption management. This method is based on supervised online learning of system state, and uses a small scalable neural network to close and open the link, so as to improve the prediction ability. Based on artificial neural network, the model makes use of the very low hardware resources and can be integrated in the large mesh and the torus NOC. Compared with static threshold calculation, The simulation results of various comprehensive traffic models on different network topologies show that this method can save the power consumption with the low hardware overhead. It can provide the ideas for solving the power consumption problem in link management NOC.

Keywords: network on a chip; static threshold; ANN mechanism; link management

0 引言

链路管理是现代片上互连的关键要素, 片上网络 (NoC) 中的功耗问题一直是研究人员关注的问题之一^[1], 最耗电的 NoC 组件之一是连接路由器之间的链路以及片上互连网络的处理元件^[2]。这些链路可以用差分信号和低电压摆幅硬件设计, 利用电平转换器作为低功耗的电路优化, 即使在没有发送有用数据时, 这些链路也几乎一直处于活动状态, 因此在没有路由器间通信出口时, 也会消耗能量。即使传统的硬件设计技术有助于降低链路的功耗, 但随着 NoC 规模的扩大, 系统级技术对于更有效地降低功耗变得

必不可少。例如, 相关研究集中在关闭和打开链路以降低功耗^[3-4], 因为系统中的某些链路在特定的运行时间框架内未得到充分利用。这些开/关动态电源管理工作大多数采用传统的链路利用率静态阈值计算, 并基于该阈值, 链路关闭一段时间, 然后在算法决定时重新打开。文献 [5] 提出了使用控制理论来管理关闭候选链路的方法, 与基于静态阈值的方法相比, 取得了良好的效果。

在文献 [5] 研究成果的基础上, 本文提出了利用应用流量的信息, 将人工神经网络 (ANN, artificial neural network) 作为动态链路功耗管理的机制。基于其动态接受

收稿日期: 2021-08-25; 修回日期: 2021-09-23。

基金项目: 北京理工大学横向科研项目 (2020I032)。

作者简介: 许威 (1982-), 男, 北京人, 硕士, 高级工程师, 主要从事电气设计及自动化方向的研究。

引用格式: 许威, 张霞. 基于人工神经网络的 NoC 智能动态链路管理方法[J]. 计算机测量与控制, 2022, 30(3): 168-172.

可变场景训练的能力,ANN可以提供灵活性和高预测能力。ANN以离散时间间隔接收链路利用率数据,并根据计算出的阈值预测应该关闭或打开的链路;同时,ANN可以动态地训练到新的应用程序信息,并且已经证明可以在类似的场景中提供准确的预测结果^[6]。此外,通过对NoC进行分区,可以分配单个小的ANN来独立监控每个分区,并行地监控整个网络^[7]。以期为解决链路管理NoC中的功耗问题提供新思路。

1 基于ANN的阈值计算方法

1.1 开/关键路的静态阈值计算

实现所提出的ANN方法的第一步是算法建立,用于比较智能管理与非智能管理是否具有可比性,不仅在节能方面,而且在吞吐量和硬件开支方面。因此,使用NoC仿真算法测试了一个案例,其中通过静态阈值机制来确定链路是关闭还是重新打开。该算法以实验NoC系统中所有链路的利用率为输入,根据静态阈值输出控制信号;对于阈值,该算法决定哪些链路被关闭。静态定义的阈值是基于不同合成流量模型的仿真观察以及与没有该机制的系统相比所观察到的功耗节省和吞吐量减少来计算的^[8]。图1所示为在4×4 NoC上观察到的4种合成流量模型的实时节能效果。

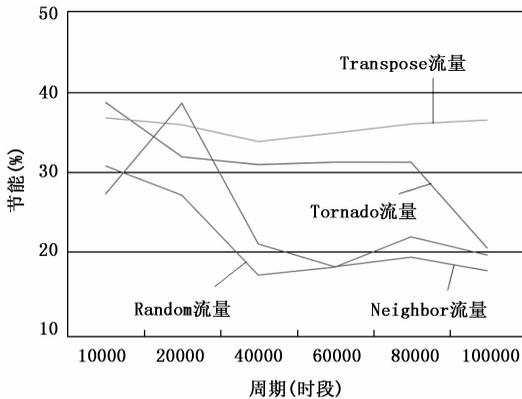


图1 4种合成流量模型的实时节能效果

1.2 ANN机制

基于ANN的机制可以集成为NoC中的独立处理元件,每个基本ANN机制可以被分配来监控NoC分区,如图2所示。每个基本ANN机制监控其区域内的所有平均链路利用率,在每个间隔期间计算其区域内每个链路的阈值利用率;然后使用该阈值来关闭该区域中表现出较低利用率的任何链路,已关闭的链路在一段时间内保持关闭状态。

提出的ANN机制使用100个周期间隔,在此期间接收所有新的利用率,这100个周期的时间间隔具有更好的节能性能。在间隔期间,ANN计算并输出新的阈值,然后由每个路由器中的链路控制机制使用该阈值来关闭未充分利用的链路。在100周期间隔期间,关闭的链路不参与下一阈值的计算;相反,用一个sentinel值进行编码,该值表示

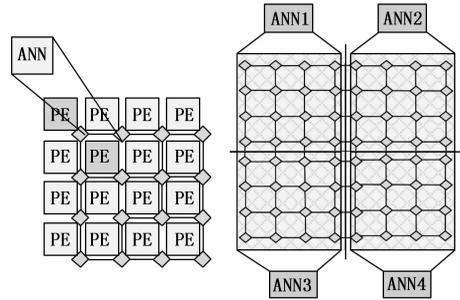


图2 ANN预测器及结构

充分利用,因此该链路不会在随后的两个时间间隔中被关闭,这将在网络中保留公平的路径分配。每个基于ANN的机制遵循一个完全连接的多层感知器模型^[9],由一个内部神经元节点的隐藏层和一个输出层神经元组成。在这项工作中,使用的激活函数是双曲正切函数,它是对称和渐近的,因此很容易在硬件中作为LUT实现^[10]。此外,该函数已在多个人工神经网络中得到广泛应用,其精度非常高^[11-13]。输入神经元的数目取决于系统接收到的反馈链路的数目,因此,ANN的大小取决于系统输入的数量。输出神经元选择与通过隐藏层神经元观察到的模式最匹配的阈值,并将阈值输出给链路控制器。

神经元计算主要为链路利用率输入加权求和的计算^[14],然后将激活函数应用于神经元输入的加权求和,以产生神经元输出(即激活神经元)。式(1)表示神经元输出计算:

$$f(x) = K \left(\sum_i w_i g_i(x) \right) \quad (1)$$

其中:K表示双曲正切的激活函数, w_i 表示应用于 $g_i(x)$ 输入函数表示的链路利用率输入的权值。整个过程如图3所示。

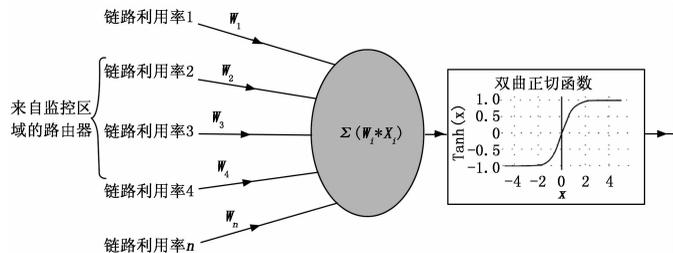


图3 神经元计算

1.3 人工神经网络的训练和操作

训练阶段可以离线执行,即在不使用NoC的情况下,并且训练权重可以存储在基于SRAM的LUT中,用于快速和在线地重新配置网络。网络是使用应用流量模式进行训练的,离线时,使用反向传播ANN训练算法^[15]。在实验中,使用了综合流量模式和Matlab神经网络工具箱;然后将权重值作为输入输入到模拟器中,在模拟器中执行和模拟实际预测。通过对系统实际运行的应用程序进行分类,可以潜在地改进ANN操作。因此,对于每个应用类别,可以使用相应的权重来训练ANN^[16]。然后,可以在

长操作间隔期间动态加载每个训练集，系统将迁移到新的应用程序。

1.4 智能阈值计算

虽然 ANN 在基于学习算法的场景预测方面非常有效，但需要仔细考虑硬件设计，因为 ANN 的大小和复杂性取决于接收到的输入数量以及必须执行的不同输出预测的数量。NoC 由大量的链路组成，这些链路随着 NoC 规模的增长呈指数级增长^[17]。因此，接收链路利用率并确定控制哪些链路可关闭和打开的阈值需要指数级可伸缩的 ANN。本研究设计了一种预处理技术，该技术基于模拟和观察来识别用于关闭和打开的候选链路集，消除几乎总是使用的链路。通过各种合成流量模拟，对于每个给定的 NoC 拓扑，计算出每个链路在模拟中各个阶段的平均利用率值^[18]，并且开始假设利用率最高的链路将处于开启状态。显然，这一步骤会降低神经网络的有效性，但有必要在性能和硬件资源方面使神经网络的规模和支出最小化。然而，确定 ANN 使用的链路以及 ANN 训练两个步骤都可以在 NoC 设计阶段离线完成。当新的应用知识可用时，神经网络训练也可以重复进行，这可能会改变片上网络流量行为^[19]。ANN 的这一特殊特性相对于静态计算阈值提供了一个比较优势，使 NoC 具有灵活性。

此外，可以将大型 NoC 划分为更小的区域。因此，可以分配一个基础 ANN 体系结构来监视每个区域，NoC 分区路由器的所有链路利用率到达负责该区域的 ANN^[20]。然而，NoC 区域的大小取决于两个主要因素：基础 ANN 提供的节能效果以及由此产生的 ANN 规模和硬件支出，这些支出随着 NoC 区域的增长呈指数级增长。

1.5 基础 (4×4) ANN 操作

ANN 机制负责计算每个时间间隔内所有链路利用率的最小值。基于最小值，神经网络计算出一个最佳阈值。图 5 所示为 4×4 NoC 分区 ANN 机制的主要过程。ANN 机制接收来自 4×4 NoC 分区的所有链路的平均链路利用率。这些值被输入到 ANN 中，以计算最佳阈值。每个路由器包含一个控制硬件监视器，用于测量每个路由器的 4 个链路中每个链路的平均链路利用率，并且该值每 n 个周期发送到 ANN。如果一个路由器不能在一个单一的时间间隔内传输这些值，则它的值被设置为 sentinel 值，这表明它的缓冲区被充分利用。该机制也可以作为拥塞信息机制，因为高度活跃的链路不是要关闭的候选链路。ANN 使用利用率值来寻找阈值，该阈值将确定在下一个 n 周期间隔内链路是关闭还是打开。如前所述，在模拟中使用了 100 个周期间隔 ($n=100$)。

1.6 基础 (4×4) ANN 硬件结构

ANN 的主要优点之一是当神经元数量很少且激活函数比较简单时，其硬件实现相对简单。由于可以将神经元运算建模为乘法累加运算，因此可以在硬件中有效地设计神经元运算。神经网络硬件实现依赖于隐藏层神经元的数

量，每个神经元都被实现为乘法累加器 (MAC) 单元，每个神经元的累加器被复用，从而使乘法器的数量最小化。基本的 ANN 硬件架构如图 4 所示。每个链路的利用率值都会通过输入协调单元来到达并排序，该单元将值分配给每个适当的乘法器。乘法器接收这些值并通过共享权重存储器获得相应的权重。然后，将权值和输入乘积累积到相应的累加器中，整个过程由有限状态机控制器控制。每个神经元都有一个指定的存储寄存器，以实现数据重用；当计算一层神经元时，它们的输出被存储在相应的寄存器中。因此，相同的硬件被重用，可用于计算下一层（即从输入层到隐藏层，从隐藏层到输出层）。当每个神经元完成 MAC 计算后，通过激活函数 LUT 计算结果并传输到输出神经元。

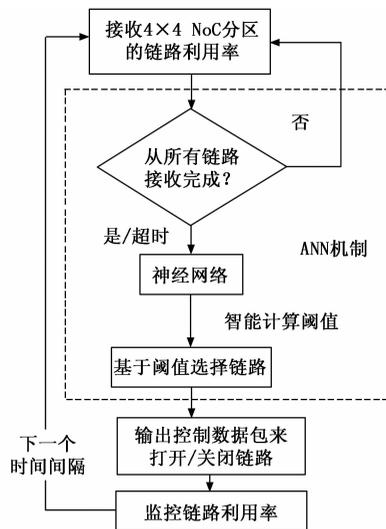


图 4 4×4 神经网络预测器的主要步骤

例如，在环面拓扑中监控 4×4 区域的 ANN 接收 64 个不同的输入，如果每个路由器在每个时间间隔内发送一个具有自己链路利用率的数据包，那么在每个周期内，ANN 最多会收到 4 个输入值。因此，如果使用流水线乘法器，则每个人工神经网络只需要 4 个乘法器就可以实现最大吞吐量。无论监控的网络大小如何，ANN 仍然是小的和灵活的。此外，监控 4×4 NoC 分区的 ANN 接收 16 个数据包（每个路由器一个），因此，需要 $16m$ 个周期（其中 m 为每个乘法器的周期延迟），加上每个累加器的 16 个周期以及激活函数的一个周期、输出神经元的一个周期，输出新的阈值，总共 $16m+18$ 个周期。

1.7 ANN 硬件优化和权衡

为了使神经网络结构更简单、更小，研究了隐藏层神经元的数量对系统总功耗的影响。鉴于 4×4 ANN 监控 16 个路由器，因此至少需要 8 个输入神经元。在神经网络的输入层有 8 个神经元意味着隐藏层应该有 5 个神经元。3 个不同的人工神经网络分别在隐藏层有 5 个、4 个和 3 个神经元。图 5 所示为在 4 种不同流量模式 (Random, Tornado,

Transpose, Neighbor) 下, ANN 的节能效果。由该图可知, 在隐藏层中使用 4 个神经元可以为所有的流量模式提供最佳的功耗节省。此外, 本文研究了训练权重的比特表示对阈值计算以及总功耗节省效果的影响, 如图 6 所示。由该图可知, 24 bits、16 bits、8 bits 和 6 bits 显示出类似的功耗节省, 但是当使用 4 bits 时, 由于降低了训练精度, 因此功耗节省效果显著降低。基于以上分析, 从 6 bits 中选择权重位表示, 使得乘法累加器硬件非常小, 每一个权重需要 6 位端口, 利用率需要 5 位端口。

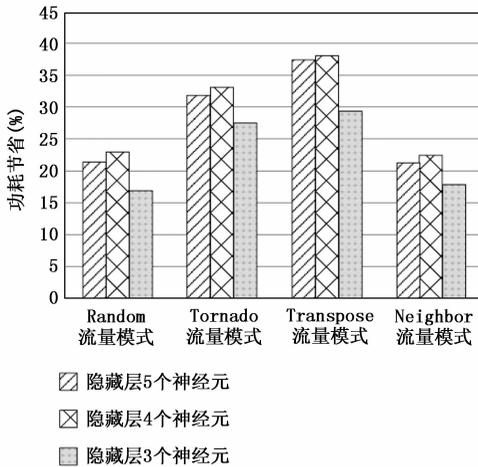


图 5 神经网络隐藏层中 5 个、4 个和 3 个神经元的功耗节省

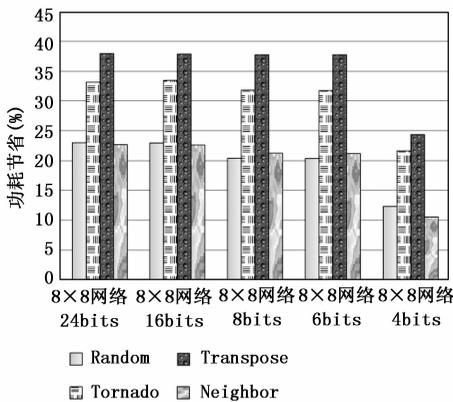


图 6 不同训练权重位表示的功耗节省

2 仿真测试

2.1 实验设置

为了评估基于链路预测机制的神经网络模型, 开发了一个基于 Java 的通用片上网络结构模拟器仿真框架。该框架利用 XY 维序路由算法, 支持虚拟信道和 4 级流水线路由器操作, 实现了多拓扑的仿真。模拟路由器支持 64 位移动宽度, 每个链路支持 2 个虚拟通道, 每个虚拟通道支持 2 个缓冲区, 该框架支持各种合成和用户定义的流量模型。

本研究实验了 4×4 网状拓扑以及 8×8 网状和环面拓

扑。在 8×8 拓扑中, 将 NoC 划分为 4×4 路由器链路的 4 个区域, 其中每个基于 ANN 的模型负责监控。基于神经网络的模型监控了对应分区中的所有链路, 所有的链路都是 off/on 的候选链路, 并基于这些体系结构细节给出了与神经网络的规模和操作相关的所有神经网络结果。

2.2 实验过程

首先, 将时间划分为 100 个周期间隔, 在每个间隔结束时, NoC 分区中的所有路由器都传输该范围内的平均利用率数据。为了保持合理的时延, 使用了一种与每个路由器对神经网络机制的期望延迟相等的超时机制。ANN 从每个路由器接收一个包, 其中每个端口有 4 个利用率值。然后, 神经网络继续计算通过控制分组传输到每个路由器的新阈值。每个路由器关闭每个链路, 具体取决于其利用率值是否高于或低于新阈值。路由器继续运行, 直到新的时间间隔结束。当链路关闭或打开时, 将额外的 100 个周期惩罚插入模拟中, 以表明对网络吞吐量的影响。

2.3 仿真结果与讨论

本文使用具有不同注入速率的合成流量模式 (Random、Tornado、Transpose 和 Neighbor), 首先评估了与没有任何开/关链路能力的相同系统和采用静态确定阈值的系统相比, 基于 ANN 机制的功耗节省。然而, 为了计算环面网络中的功耗节省, 还测量了环面网络所有可能分区中的链路利用率, 以补偿环面链路, 具有最小值的链路利用率通过 ANN 传递。

图 7 所示为基于 ANN 技术、静态阈值技术的 8×8 网状和 8×8 环面拓扑的节能效果。由该图可知, 基于 ANN 机制比传统的静态阈值机制节省了更多的流量。基于 ANN 机制可以在流量模式中识别出大量的未来行为, 因此, 它可以智能地选择下一个定时间间隔所需的阈值。为了得到所提出方案的节能效果, 针对商用 65 nm CMOS 技术合成并实现了一个 8×8 NoC 区域的基于人工神经网络的机制。在每个路由器中, 基于 ANN 的控制器和相关的硬件开销消耗了大约 4 k 个逻辑门, 从而将 8×8 网状的估计硬件开销提高到 NoC 硬件的 4% 左右。

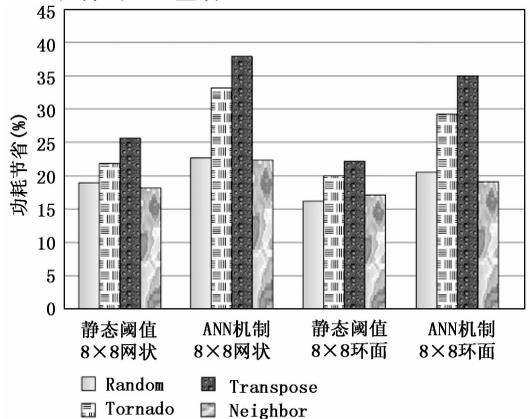


图 7 8×8 网状和环面拓扑的节能效果

图 8 所示为 8×8 网状和环面拓扑的吞吐量比较。由该图可知，虽然没有开/关机制的吞吐量高，但与静态确定阈值技术相比，基于 ANN 的技术显示出更好的吞吐量结果。这取决于所选择的网络拓扑（例如，在 2D 网格拓扑中，可能不繁忙的链路包括位于网格边缘的链路，而中心链路通常更活跃并且可以一直保持打开），因此 ANN 机制可以以更易于管理的方式处理输出决策，这可以通过智能器件的规划和 NoC 内部处理元件的放置来实现。

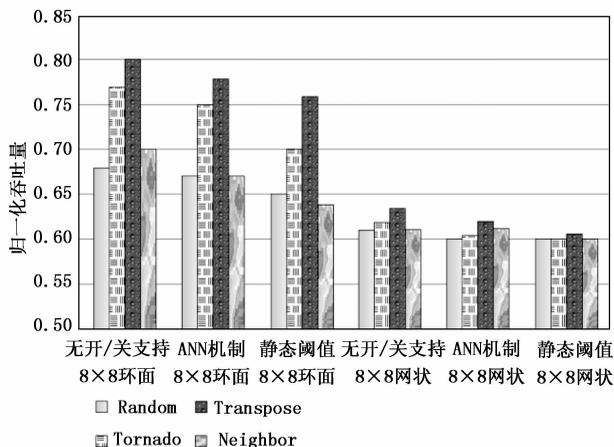


图 8 8×8 网状和环面拓扑的归一化吞吐量比较

图 9 所示为 8×8 网状网络的归一化能耗。可以观察到，使用 ANN 机制的功耗比静态计算阈值和不使用开关链路管理算法的情况要少。与没有开/关链路或与静态阈值相比，ANN 表现出整体功耗的降低。这是因为基于人工神经网络阈值计算方案具有动态、智能、灵活的特点，可以随时确定哪些链路可以关闭或打开。

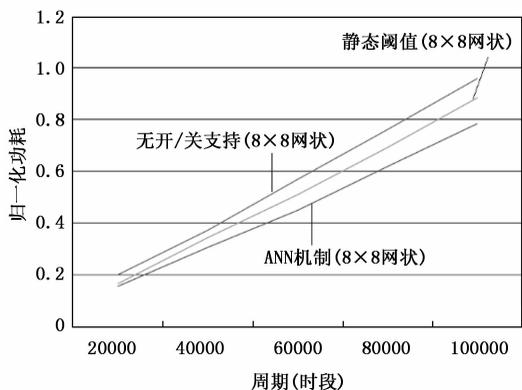


图 9 8×8 网状网络的能量消耗

3 结束语

本文提出了一种基于人工神经网络机制来动态计算链路利用率阈值，进而使用利用率阈值来选择开启或关闭候选链路，以达到节省 NoC 功耗的目的。基于人工神经网络的模型利用了非常低的硬件资源，并且可以集成在大型网

状和环面 NoC 中，显示出显著的功耗节省效果。仿真结果表明，在综合流量模型下，与静态阈值方法相比，可节省功耗。在未来的研究中，还需进一步探索利用人工神经网络等智能方法动态降低 NoC 功耗的方法。

参考文献:

- [1] 杨晓燕. 基于 NoC 模拟器的片上网络影响因素研究 [J]. 九江学院学报 (自然科学版), 2019 (4): 63-65.
- [2] 孙晓乐, 钱亚龙, 齐新新, 等. 片上互连网络的功耗特征分析与优化 [J]. 计算机工程与科学, 2020 (7): 1141-1150.
- [3] 刘玉婷, 宋荣方. OFDM-NOMA 下行链路用户分组及功率分配研究 [J]. 南京邮电大学学报 (自然科学版), 2018, 38 (5): 32-38.
- [4] 赵通. 面向物联网应用的低功耗 SoC 平台设计 [D]. 杭州: 浙江大学, 2018.
- [5] 孙晓乐, 钱亚龙, 齐新新, 等. 片上互连网络的功耗特征分析与优化 [J]. 计算机工程与科学, 2020 (7): 1141-1150.
- [6] 梁浩, 黄生志, 孟二浩, 等. 基于多种混合模型的径流预测研究 [J]. 水利学报, 2020, 51 (1): 112-125.
- [7] 娄雷雷. 面向 NoC 众核容错系统的动态任务分配机制设计 [D]. 哈尔滨: 哈尔滨工业大学, 2019.
- [8] 邹承明, 刘攀文, 唐星. 动态主机配置协议泛洪攻击在软件定义网络中的实时防御 [J]. 计算机应用, 2019, 39 (4): 1066-1072.
- [9] 袁光耀. 基于非线性滤波优化的前馈神经网络训练方法研究 [D]. 开封: 河南大学, 2016.
- [10] 苏潮阳, 应三丛. 一种神经网络的可配置激活函数模块设计 [J]. 单片机与嵌入式系统应用, 2020, 20 (4): 6-9.
- [11] 常利苹, 曹飞龙. 一类神经网络算子的构造与逼近 [J]. 中国计量大学学报, 2019, 30 (3): 337-342.
- [12] 石建飞, 衣淑娟. 输入有界的电液位置伺服系统的自适应控制 [J]. 系统仿真学报, 2017, 29 (3): 676-682.
- [13] 李子富. 非线性随机系统自适应神经网络控制 [D]. 大连: 大连海事大学, 2015.
- [14] 杨程炜. 基于节点重构思想的链路预测算法研究 [D]. 武汉: 武汉大学, 2019.
- [15] 张焕. SDN 环境下基于神经网络的网络流量预测模型研究 [D]. 长沙: 湖南师范大学, 2019.
- [16] 刘琪, 温晓. 基于人工神经网络的 WENO 重构方法 [J]. 应用数学进展, 2020, 9 (4): 574-583.
- [17] 金世燕, 姚远程, 秦明伟. 基于通信链路的 NoC 映射算法 [J]. 电子技术应用, 2016, 42 (8): 121-124.
- [18] 曹景奇. 面向有线/无线混合 NoC 的低功耗路由算法研究 [D]. 哈尔滨: 哈尔滨工程大学, 2016.
- [19] 翁凯衡, 韩银和, 王颖, 等. 一种基于 3D 内存用于神经网络加速的可数据压缩 NoC [C] // 中国计算机学会, 2017.
- [20] 姜奎, 韩国栋, 沈剑良. 一种基于分层 Mesh 网络的层次化 NoC 拓扑结构 [J]. 计算机工程, 2017, 43 (10): 1-5.