面向智能路由的光子脉冲强化学习

Shuiying Xi ang[1]*, Yonghang Chen[1], Ling Zheng[1,2]*, Zhicong Tu[1], Xintao Zeng[1], Mengting Yu[1] ,

Shuai Wang[1], Yahui Zhang[1], Xingxing Guo[1], Weitao Pan[1]*, Yue Hao[1]

[1]综合业务网国家重点实验室，西安电子科技大学，西安 710071，中国；[2]西安邮电大学通信与信息工程学院，西安 710121，中国*通讯作者：syxiang@xidian.edu.cn，lingzheng@xupt.edu.cn， wtpan@mail.xidian.edu.cn

摘要：智能路由在现代通信基础设施中发挥着关键作用，应用场景包括数据中心、计算网络和未来的6G网络。尽管强化学习（RL）在智能路由方面展现出巨大潜力，但其实际部署仍受高能耗和决策时延的限制。本文提出一种光子脉冲强化学习架构，用于实现基于近端策略优化（近端策略优化）的智能路由算法。我们在具有胖树拓扑的软件定义网络（软件定义网络）上对所提出方法的性能进行了系统评估。结果表明，在不同基准流量速率条件下，基于近端策略优化的路由策略在吞吐量、丢包率、平均时延和负载均衡等关键性能指标上显著优于传统的Dijkstra算法。此外，我们针对三种典型基准流量速率，实现了脉冲Actor网络的软硬件协同框架，其中采用了基于马赫-曾德尔干涉仪（MZI）阵列的光子突触芯片，以及基于带可饱和吸收体的分布反馈激光器（分布反馈激光器-可饱和吸收体）的光子脉冲神经元芯片。针对640个状态-动作对的实验验证表明，软硬件协同框架的推理准确率与纯算法实现一致。本文进一步分析了脉冲Actor网络中不同隐藏层规模以及胖树拓扑网络规模变化的影响。光子脉冲强化学习与基于软件定义网络的路由相结合，为智能路由优化建立了一种新范式，具备超低时延和高能效的特点。该方法在实时网络优化场景中展现出广阔的应用前景，包括大规模数据中心、计算网络、卫星互联网系统和未来的6G网络。

关键词：光子脉冲神经网络；脉冲强化学习；智能路由；软件定义网络

1. 引言

云计算、大数据、人工智能和物联网等新兴信息技术，正以前所未有的速度重塑社会生产和生活方式。这一变革伴随着数据量的爆炸式增长，以及日益严格的实时处理需求，对底层网络基础设施的性能、效率和智能化水平提出了极高的

要求。路由技术负责资源分配优化、服务质量保障和整体效率提升，已成为信息通信技术领域的重要研究前沿[1]。

以基于最短路径的Dijkstra算法及其在开放最短路径优先（OSPF）等协议中的实现为代表的传统路由算法，长期以来一直是互联网路由的基础[2]。然而，这些方法依赖于由局部或固定度量（例如跳数和链路带宽）驱动的静态或准静态确定性决策机制。面对现代网络中高度动态、突发且异构的流量模式，尤其是在数据中心和算力网络中，这一范式存在固有局限。首先，它们缺乏实时全局状态感知和历史学习能力。其次，其固定规则策略过于僵化，难以自适应地学习并应对未知或复杂的非线性流量模式。最后，其单目标优化（例如最小化跳数）往往导致局部拥塞和资源利用效率低下。

为克服传统路由方法的局限性，深度强化学习（DRL）已成为智能路由优化的一种颇具前景的范式[3-7]。深度强化学习智能体能够从高维状态空间中学习复杂的动态特性，并通过与网络环境的持续交互，收敛到全局优化的路由策略[8-12]。在各类深度强化学习算法中，近端策略优化（PPO）算法凭借其训练稳定性、较高的样本效率以及对连续和离散动作空间的支持，在复杂路由优化任务中展现出强大潜力[13]。然而，采用传统人工神经网络（ANNs）的基于深度强化学习的路由仍然面临较高的计算能耗和决策时延，这与实际网络控制平面对超低时延和资源受限的要求相冲突——尤其是在边缘计算场景中。

受生物神经信息处理启发，脉冲神经网络（SNNs）为提升能效和时间动态特性提供了一种颇具吸引力的替代方案，相较于基于人工神经网络的模型更具优势 [14，15]。与此同时，光子计算利用光的固有物理优势，包括高速、大规模

并行性、超宽带宽和高能效，从而为神经形态计算提供强大的硬件平台 [16, 17]。通过将脉冲神经网络的计算原理与光子硬件相结合，光子脉冲神经网络（PSNNs）已成为构建下一代高性能、高能效智能计算架构的关键方向 [18, 19]。基于带可饱和吸收体的分布反馈激光器（DFB-SAs）芯片的 PSNNs，已在图像分类等代表性任务中取得初步成果 [20, 21]。近年来，光子芯片已拓展至强化学习相关应用领域 [22, 23]。Jin等人开发了一种可编程混合光子芯片，能够以高精度执行光学点积运算，维度最高可达 15，从而能够在光子平台上运行复杂的强化学习算法。然而，将这项新兴技术应用于智能路由这一关键任务的研究仍鲜有涉及。

在本研究中，我们提出了一种基于光子脉冲强化学习的软件定义网络（SDN）智能路由架构[24, 25], ，该架构将 PPO 算法强大的学习与优化能力，与 PSNN硬件的超低时延和高能效紧密结合。本研究旨在弥合光子神经形态计算与网络路由技术交叉领域存在的研究空白。本研究的主要贡献概括如下：

（1） 架构与算法创新。我们开发了一种基于 PPO 框架的智能路由模型，其中，负责路由决策的策略网络由脉冲神经网络实现。该设计使路由智能体的计算范式与光子神经形态硬件的特性相契合，从而能够高效地部署于硬件。我们将所提出的模型集成到软件定义网络控制平面中，并在具有代表性的胖树数据中心拓扑上对其进行评估。

(2) 性能提升。在仿真的胖树网络中开展了广泛的对比实验，以评估所提出的光子脉冲PPO路由算法与传统 Dijkstra 算法的性能。通过对吞吐量、时延、丢包率和负载均衡进行全面评估，持续验证了我们方法卓越的整体性能。

(3) 软硬件协同计算。我们开发了一个软硬件协同计算框架，用于将脉冲Actor网络映射至 PSNN硬件。在 640 个具有代表性的状态 – 动作对上进行的实验验证表明，从软件仿真映射到硬件执行的过程中，决策准确率和算法性能均得到了完整保留。

本文其余部分安排如下。第二节介绍所提出的基于光子脉冲强化学习的智能路由框架的整体架构和运行原理，描述用于软硬件协同计算的实验平台，并系统阐述光子脉冲PPO路由算法的设计、实现与训练方法学。第三节展示并分析在胖树网络拓扑下获得的仿真结果。第四节报告软硬件协同计算结果。第五节研究不同SNN隐藏层规模及不同胖树网络规模的影响。最后，第六节总结主要研究发现。本研究建立了一种全新的智能路由优化范式，兼具超低时延、高能效和更强智能性，对未来的大规模数据中心、计算网络、卫星互联网系统及6G网络具有重要价值。

2. 用于智能路由的光子脉冲 PPO 整体架构

2.1 整体架构

所提出的基于光子脉冲强化学习的 SDN智能路由系统采用分层架构，由数据平面、控制平面和应用平面（智能决策层）组成，如图 1 所示。数据平面采用参数为 k = 4 的胖树拓扑，由多个支持OpenFlow的可编程交换机组成。它负责实时数据包转发，并持续采集网络状态信息。包括链路带宽、时延和负载条件在内的关键指标会被采集，并通过OpenFlow协议报告给控制平面。同时，数据平面通过安装控制平面下发的流表规则来执行路由决策。本研究中的 Fat-tree 网络包含 16 台主机和 20 台交换机，链路容量统一为 1000 兆比特每秒。训练期间，基准流量速率设为 50 兆比特每秒，且

流量对固定为 64。为进行性能评估，基准流量速率从 40 兆比特每秒变化至 200 兆比特每秒，以全面评估不同网络负载条件下的路由性能。

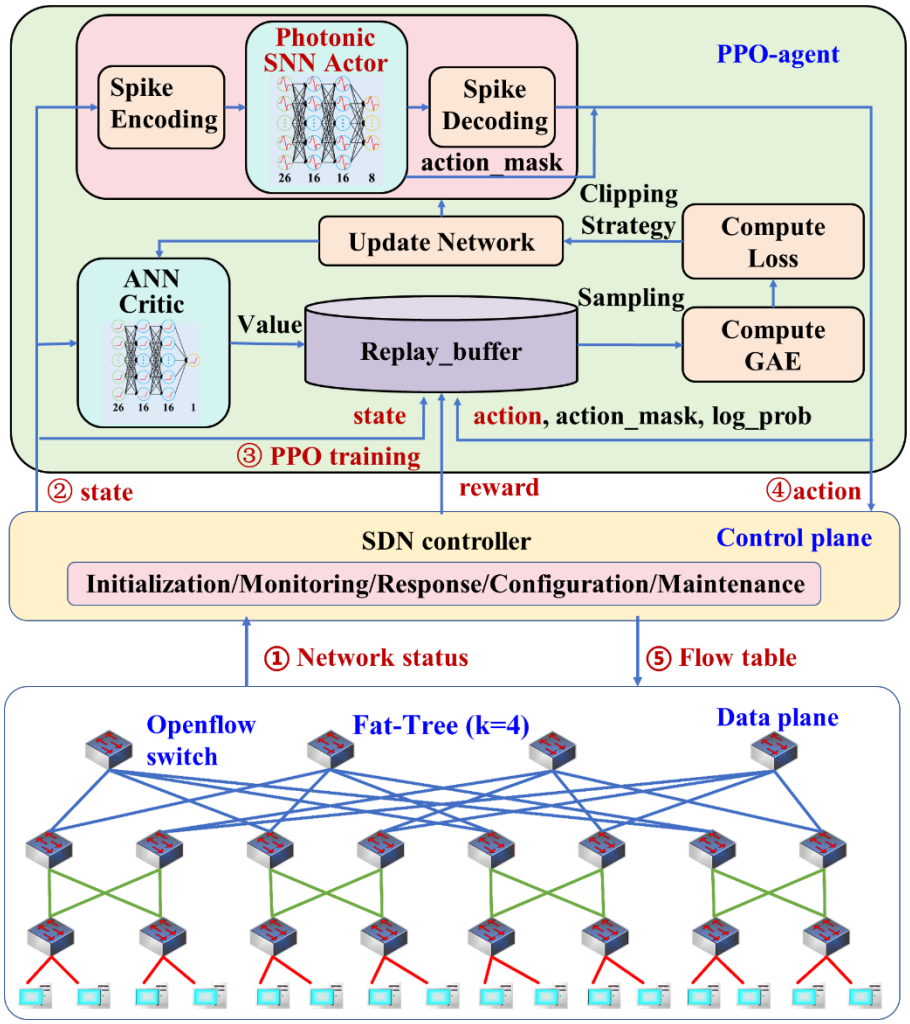

图 1 用于 SDN 智能路由的光子脉冲 PPO 总体架构。

控制平面构建于 SDN控制器 之上，提供集中式网络管理和路由执行。它由五个功能模块组成：**初始化**，用于注册交换机并建立控制信道；**监控**，用于定期收集网络状态信息（例如链路利用率和数据包统计信息）；**响应**，用于处理 Packet-In 消息并触发路由决策；**配置**，用于安装由智能决策层生成的流表规则；以及**维护**，用于管理流表超时和网络状态更新。

应用平面是系统的核心创新，采用基于PPO 演员—评论家框架的光子脉冲强化学习引擎实现。它处理网络状态信息并

生成优化的路由策略。光子脉冲演员网络通过脉冲神经网络实现，由三个模块组成：脉冲编码、脉冲神经网络层和脉冲解码。网络状态被编码为脉冲序列，由脉冲神经网络处理以生成路由动作，再通过动作掩码进行解码，过滤无效决策。Critic网络采用人工神经网络实现，用于评估状态价值并辅助策略优化。PPO训练模块包含经验回放缓冲区、广义优势估计（GAE）、损失计算、裁剪策略和网络更新机制，从而实现稳定高效的策略优化。

脉冲Actor网络采用受生物神经动力学启发的漏积分发放（LIF）神经元作为基本计算单元。在每个离散时间步t，神经元的膜电位u_t根据式（1）和式（2）更新：

$$u_{t+1} = \tau \cdot u_t + x_t - V_{th} \cdot S(u_t) \tag{1}$$

$$S(u_t) = \begin{cases} 1, & \text{if } u_t \geq V_{th}; \\ 0, & \text{if } u_t < V_{th} \end{cases} \tag{2}$$

其中，输入电流（或输入信号）为x_t，脉冲发放阈值为V_{th}，泄漏系数τ (0<τ<1)控制膜电位的时间记忆；当电位超过预设阈值时，神经元会发放脉冲，随后进行膜电位重置。为解决脉冲发放函数$S(u_t)$的不可微问题，采用零膨胀发放（ZIF）替代梯度方法近似脉冲激活函数，从而实现式（3）所定义的基于梯度的反向传播。此外，如式（4）所述，在训练过程中引入梯度裁剪，以稳定膜电位更新并确保脉冲神经网络的有效优化。

$$S(u_t) \approx \sigma(u_t) = \begin{cases} 1, & \text{if } u_t > 0; \\ 0, & \text{if } u_t \leq 0 \end{cases} \tag{3}$$

$$\frac{\partial \sigma(u_t)}{\partial u_t} \approx clip(-1, +1) \tag{4}$$

2.2 智能路由优化过程

基于所提出的架构，智能路由优化过程分为五个依次进行的

阶段如下：

阶段1：网络 状态采集。 数据平面中支持OpenFlow的交换机持续监测网络状况，包括拓扑结构、链路利用率、时延、流量分布和丢包等。所采集的信息（1网络状态）通过南向接口上报给控制平面中的SDN控制器。

阶段 2：状态感知与表示。 SDN控制器对收集到的网络信息进行归一化和聚合，形成适用于强化学习的状态表示(2state)，随后将其传递至PPO智能体的Actor网络和Critic网络。

阶段 3：PPO训练与策略优化。 该阶段构成光子脉冲强化学习框架的核心优化过程(3PPO training)。首先，Spike Encoding模块将输入状态转换为脉冲序列，这些脉冲序列由包含多个全连接层的脉冲Actor网络（26×16×16×8）进行处理。经过脉冲解码后，Actor输出路由动作及其对应的对数概率，同时输出动作掩码以消除无效动作。与此同时，Critic网络评估输入状态，并输出其价值以进行优势估计。元组{state, action, action mask, action probability, reward}存储在经验回放缓冲区中。在训练过程中，通过采样经验计算广义优势估计，并使用带有裁剪策略的PPO损失函数迭代更新Actor网络和Critic网络。

阶段 4：动作决策与流表生成。收敛后，训练完成的 Actor网络根据当前网络状态生成最优路由动作(4action)，并将其传输至控制平面的 SDN控制器。

阶段 5：流表下发与路由执行。SDN控制器将路由决策转换为流表项，并将其安装到数据平面的 OpenFlow 交换机上 (5Flow table)。交换机依据更新后的流表执行数据包转发，从而完成闭环智能路由过程。

2.3 软硬件协同训练推理框架

光子硬件计算本质上容易受到制造缺陷和系统噪声的影响，可能导致精度下降。为缓解脉冲神经网络训练的挑战以及光学推理误差，我们采用一种

端到端硬件与软件协同训练–PSNN 的推理框架，遵循类似的方法学

见参考文献 [26]。

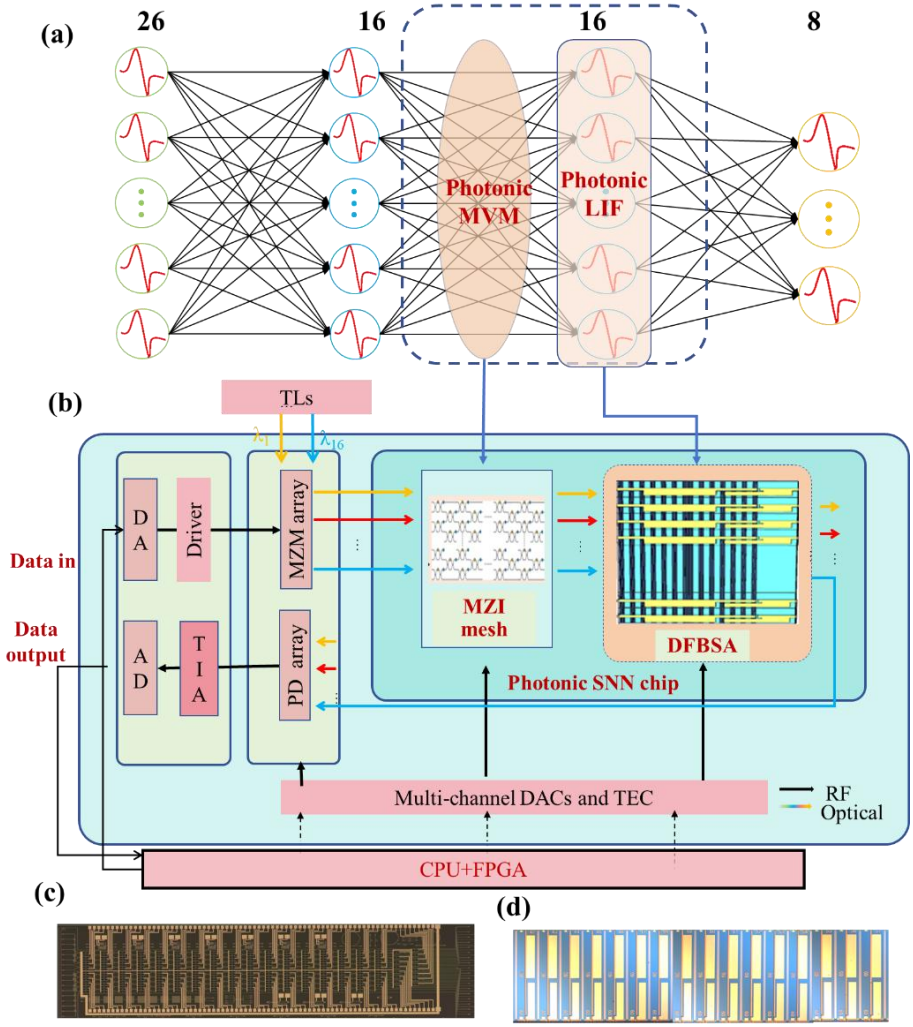

图2 光电协同计算装置示意图。(a) 脉冲神经网络架构；(b) 脉冲神经网络的光子硬件实现；(c) MZI 网格

步骤 1：软件预训练。执行者网络采用从头开始训练的脉冲神经网络，并使用单时间步（T=1）代理梯度方法，而 Critic 则由人工神经网络实现。在基于Actor–Critic的深度强化学习算法中，Critic网络不参与硬件推理，因此仍采用人工神经网络实现，从而能够准确估算折扣累积奖励，并实现稳定的策略学习。

步骤 2：MZI 芯片的光子硬件原位训练。通过在线训练，将预训练脉冲Actor网络中隐藏层线性变换的 16×16 权重矩阵导出并映射到 MZI 网格芯片上。为了将数字权重映射到光子硬件上，采用随机并行梯度下降（SPGD）

算法来优化 MZI 网格的移相器加热器电压[27]。基于 SPGD 的训练框架会迭代执行随机电压扰动、梯度估计和电压更新。在仿真中，使用步骤 1 得到的收敛隐藏层权重矩阵作为目标矩阵。通过迭代进行 SPGD 优化，可获得 MZI 移相器的电压配置表，进而得到相应的硬件等效权重矩阵，完成从数字域权重到光学域实现的映射。

步骤 3：硬件感知软件微调。 MZI 训练完成后，将得到的权重矩阵重新加载到算法中进行再训练，以补偿制造误差和噪声引起的精度损失。固定中间线性层权重，并禁用梯度。再训练后，导出相关层的输入和输出，用于软硬件协同推理。在三种负载条件（40、80 和 120 兆比特每秒）下，使用 640 × 16 状态–动作样本对进行硬件验证。

步骤 4：硬件感知协同推理。光电协同计算框架如图 2(b) 所示。受当前 PSNN 芯片规模限制，仅将中间隐藏层部署在光子平台上，包括光子矩阵–向量乘法（MVM）和光子 LIF 神经元模块。具体而言，先前 26 16 网络的非线性脉冲激活输出由 CPU–FPGA 平台计算，并用作光子计算网络的输入，再传输至数模（DA）转换器和驱动电路。这些电子信号由马赫–曾德尔调制器（MZM）阵列调制到可调谐激光器阵列产生的光载波上。随后，调制光信号由 MZI 网格处理，以执行线性矩阵–向量乘法运算，接着由 DFB-SA阵列实现非线性脉冲激活。所得光脉冲信号通过光电探测器（PDs）转换回电子域，经跨阻放大器（TIA）放大，再由模数（AD）转换器数字化，从而完成一层光电混合计算，然后输入后续层。在实验过程中，使用可编程多通道电压源和高精度温度控制，对 MZI 网格和 DFB-SA阵列进行精确驱动和稳定控制，确保光子脉冲推理可靠且可重复。记录的输出

是然后返回到算法的第三线性层，进行下一层计算。最后，将所得结果与微调软件模型的结果逐项比较，以计算软硬件协同计算准确率。

2.4 光子脉冲 PPO 算法

为实现所提出的光子脉冲 PPO 算法，构建了一个与实际软件定义网络场景一致的仿真环境，其中包括控制器、交换机和主机。该定制环境将状态定义为网络特征，将动作定义为路径选择，并将奖励定义为吞吐量、丢包率、平均时延和负载均衡的联合函数。

(1) 状态空间。状态空间旨在捕获路由决策所需的关键信息，由主机编码和路径特征组成。主机编码包括源端和目的地主机的归一化索引[src_{norm}、dst_{norm}]，使智能体能够区分不同的流量对。路径特征部分最多包含 8 条候选路径，每条路径由 3 个特征描述：平均利用率 $U_{avg}(p)$、最大利用率 $U_{max}(p)$和归一化跳数 $H_{norm}(p)$。这些特征共同构成一个 26 维状态向量：

$$s=[src_{norm},dst_{norm},U_{avg}(p_1),U_{\max}(p_1),H_{norm}(p_1),...,U_{avg}(p_8),U_{\max}(p_8),H_{norm}(p_8)] \tag{5}$$

详细定义如下：

$$\begin{cases} src_{norm}=\dfrac{index(src)}{N_{hosts}},dst_{norm}=\dfrac{index(dst)}{N_{hosts}},U_{avg}(p)=\dfrac{1}{|E_p|}\sum_{(u,v)\in E_p}, \\ \dfrac{L(u,v)}{C}U_{max(p)}=max_{(u,v)\in E_p},\dfrac{L(u,v)}{C}H_{norm}(p)=\dfrac{|P|-1}{10} \end{cases} \tag{6}$$

其中，索引(·)表示主机索引；N_{host} 为主机总数；E_p 表示路径p的边集；$L(u, v)$是链路(u, v)的当前负载；C是链路容量（1000 兆比特每秒）；$|P|$ 表示路径p的跳数（即路径p经过的链路数）。

（2）动作空间。动作空间被定义为离散的路径选择向量，$m = min$（实际路径数，8），其中有效动作的数量根据可用路径动态变化。采用动作掩码，以确保只能选择可行路径：

$$\text{mask}[i]=\begin{cases}1 & \text{if } i<m \text{ (path exists)} \\ 0 & \text{if } i\geq m \text{ (path doesn't exist)}\end{cases} \tag{7}$$

掩码为应用　　应用于logits输出，该输出属于Actor网络，从而得到有效的动作概率分布p　　$_i$:

$$\begin{cases} p_i = \dfrac{\exp(\text{logits}'[i])}{\sum_{j=0}^{m-1}\exp(\text{logits}'[j])}, \\ \text{logits}'[i] = \begin{cases} \text{logits}[i] & \text{if mask}[i]=1 \\ -\infty & \text{if mask}[i]=0 \end{cases} \end{cases} \tag{8}$$

(3) 奖励。Th　e 奖励函数共同优化传输效率、可靠性和负载均衡等多个优化目标，
cy、可靠性和负载均衡，其定义为:

$$R_{total} = R_{base} + R_{traffic} - P_{drop} - P_{load} \times W_{factor} - P_{hop} \times W_{hop} + R_{balance} \tag{9}$$

其中，R_{base} = 30.0 提供稳定的正向奖励基线；$R_{traffic}$ 鼓励提高吞吐量；P_{drop} 对丢包进行惩罚，且流量负载越高，惩罚越重；P_{load} 对链路利用率不均衡进行惩罚，惩罚权重由网络利用率因子 W_{fActor}决定；P_{hop} 通过跳数　　的奖励公式如下:

$$\begin{aligned} R = {} & 30.0 + 0.3D - \left(1 - \frac{D}{T}\right)(1.5T + 20.0) - \left(50U_{\text{avg}}^2 + 60U_{\max}^2\right) \times \left(0.8 + 1.2U_{\text{net}}\right) \\ & -2.0H \times \left(1.2 - \min\left(U_{\text{net}}, 0.8\right)\right) + 50.0B_{\text{long}} \end{aligned} \tag{10}$$

其中，D表示成功传输的流量（单位：兆比特每秒），T表示该流量的目标传输量（单位：兆比特每秒），U_{avg}表示路径p经过链路的平均利用率，$U_{\max}$表示路径p经过链路的最大利用率，U_{net}表示整个网络的全网链路利用率，H表示路径的跳数，B_{long}表示长期负载均衡度。最后，将奖励值限制在[−40.0 至 100.0] 的范围内，以防止奖励爆炸。

为评估智能路由算法的性能，采用了四项关键指标。

（1）吞吐量。吞吐量衡量单位时间内成功传输的数据量，反映了SDN网络的整体数据承载能力。它受链路带宽、转发能力和流量调度的影响，定义如下:

$$\text{Throughput} = \frac{\sum_{i=1}^{N} ap_i \times 10^6 \times \Delta t}{\Delta t \times 10^6} = \frac{\sum ap_{\text{i}} \times 10^6}{10^6}\text{Mbps} \tag{11}$$

其中，ap表示统计周期 Δt（秒）内成功传输的数据总量（比特）。

（2）数据包丢包率。丢包率用于量化统计周期内未能到达目的地的数据包数占传输数据包总数的比例，反映网络可靠性。其通常由拥塞导致，

资源不足或发生链路故障时，其计算公式为

$$\text{PacketLossRate} = \left(1 - \frac{dp}{tp}\right) \times 100\%, tp = \frac{tm \times 10^6 \times \Delta t}{1500 \times 8} \tag{12}$$

其中，dp为成功接收的数据包总数，tp为传输的数据包总数，tm

表示网络传输速率（单位：兆比特每秒）。

(3) 平均时延。平均时延表示数据包从源端传送到目的地所需端到端时间的平均值，包括传输、处理、排队和传播时延。它直接反映了SDN网络的实时性能，计算公式为

$$\text{AverageDelay} = \frac{\sum_{i=1}^{N} (pd_i \times d_i)}{tt} \text{ms} \tag{13}$$

其中，pd 为每条传输路径的时延，d为相应的成功传输流量，tt为

传输总流量，N为统计周期内的数据包数量。

(4) 负载均衡。负载均衡用于评估网络链路间流量分布的均匀程度。平均

负载L_{avg} 和负载方差 σ2 的计算公式如下：

$$L_{avg} = \frac{1}{n}\sum_{i=1}^{n} l_i, \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(l_i - L_{avg}\right)^2 \tag{14}$$

其中，n表示资源数量，l表示每个资源的负载（例如带宽利用率）。最大

方差 σ^2_{max} 随后计算如下：

$$\sigma^2{}_{max} = \frac{(n-1)}{n} \times (L_{max} - L_{avg\_max})^2 \tag{15}$$

其中，L_{max} 是单链路的最大容量（单位为兆比特每秒），而$L_{avg_max} = L_{max} / n$。因此，负载均衡

计算如下：

$$\text{LoadBalance} = 1 - \frac{\sigma^2}{\sigma^2{}_{max}} \tag{16}$$

在训练阶段，采用 Adam 优化器，并结合 StepLR 学习率调度器和梯度裁剪，以确保训练稳定性。使用 PPO 混合损失函数，其中包括 ε-裁剪策略损失、均方误差（MSE）价值损失和熵正则化项，从而平衡策略优化、价值

估计和探索势估计和经验回放缓冲区分别用于优势计算和数据存储。网络参数以小批次迭代更新，同时自适应探索率衰减策略动态平衡探索与利用，从而高效学习软件定义网络路由策略。

所有实验均在一台配备 Intel (R) Xeon (R) Platinum 8251 CPU（3.80 GHz）和 NVIDIA GeForce RTX 4090 GPU 的平台上进行，该平台运行 Ubuntu 20.04.6 LTS，支持模型训练、推理和网络仿真。环境和训练参数的详细信息汇总于表1。

表1 用于 SDN 路由的近端策略优化训练参数

类别	参数	值
环境	网络链路带宽容量	1000 兆比特每秒
	基准流量生成率	100 兆比特每秒
	最大可选路由路径数	8
	链路安全利用率阈值	0.95
	控制器-交换机交互延迟范围	(0.001, 0.002) s
	交换机流表最大容量（核心层 / 汇聚层 / 边缘层）	100 / 150 / 200
	流表项空闲超时（核心层 / 汇聚层 / 边缘层）	30 / 20/ 15 s
	默认数据包大小	1500 字节
	流量统计时间间隔	1.0 秒
训练	执行者和 Critic 网络的学习率	5e-05
	网络隐藏层维度	16
	PPO策略更新裁剪系数	0.05
	熵奖励系数	0.05
	价值损失权重系数	2
	GAE优势估计折扣因子	0.96
	长期奖励折扣因子	0.99
	训练批次大小	32
	每轮经验的更新迭代次数	15
	探索率衰减系数	0.99
	最小探索率	0.05
	梯度裁剪阈值	0.5
	完整训练回合总数	3000

光子脉冲 PPO 算法的整体训练过程总结如下方所示的伪代码

。

算法 1 用于 SDN 路由的脉冲 PPO 训练流程

要求： 环境 $\mathbb{E}$，最大回合数 T，批量大小 N，折扣因子 γ，广义优势估计系数 λ，PPO 裁剪阈值 ε_{clip}，熵系数 λ_{ent}，价值损失系数 λ_{val}，更新轮数 K，探索衰减率 λ_{exp}，最小探索率 ε_{min}

确保： 优化后的脉冲Actor网络 π^{snn}，Critic网络 V

流程 脉冲 PPO 训练
初始化 π^{snn}, V, Adam 优化器 经验回放缓冲区 $\mathbb{B}$,
探索率 $\varepsilon = 1.0$, 奖励历史记录 $\mathcal{P}$ (窗口 = 20)
对于 $t \leftarrow 0$ 到 $T-1$ **执行**
重置 $\mathbb{E}$, 获取初始状态 s, 回合_奖励 $\leftarrow 0$
当 $\mathbb{E}$ 未终止时 **执行**
$d_a \leftarrow \mathbb{E}$.获取_动作_维度(); **如果** $d_a = 0$ 或 $s = \emptyset$ **则跳出**
a，对数$_p$，v，掩码 $\leftarrow \pi^{snn}$。select_action(s，d_a，ε)
s'、r、done、info $\leftarrow \mathbb{E}$。step(a)
将（s、a、对数$_p$、v、r、$done$、掩码）存储在 $\mathbb{B}$ 中；
回合_奖励 += info.raw_奖励；$s \leftarrow s'$
结束循环
从 $\mathbb{B}$ 中提取批次；$v_{term} \leftarrow V(s_N)$
$\mathcal{A}$, $\mathcal{P} \leftarrow$ 广义优势估计 (r [illegible] , [illegible] , v_{term}, γ, λ);
$\mathcal{A} \leftarrow (\mathcal{A}$-$\mathcal{A}$.均值())/($\mathcal{A}$.标准差()+1e-8)
循环 [illegible] $\leftarrow 0$ 到$K-1$ **做**
打乱批次索引（固定随机种子）
循环 起始 $\leftarrow 0$ 到 N 步长 N **执行**
s_b, a_b, log_{pb}, $\mathcal{A}_b$, $\mathcal{P}_b$, 掩码$_b$ $\leftarrow$ 批次 [开始: 开始 + N]
$logits \leftarrow \pi^{snn}$ (s_b).masked_fill(~掩码_b, -1e9)
$v_{pred} \leftarrow V(s_b)$; $dist \leftarrow$ 分类分布($logits$)
$\log_{p\_new} \leftarrow dist.\log_{prob}(a_b)$
损失$_{policy} \leftarrow$ -最小值(exp(log_{p_new} - log_{p_b}) × $\mathcal{A}_b$;
截断(..., 1 - ε_{clip}, 1 + ε_{clip}) × $\mathcal{A}$ b).mean()
损失$_{total} \leftarrow$ 损失$_{polic}$ [illegible] $_y$ + λ_{val} × 均方误差(v_{pred}, $\mathcal{P}_b$)
- λ_{ent} × dist.熵().mean()
更新 π^{snn} / V 通过反向传播（梯度裁剪）
结束 for
结束循环
调整 λ_{exp} by $\mathcal{P}$ 的奖励增长；$\varepsilon \leftarrow \max(\varepsilon_{min}, \varepsilon \times \lambda_{exp})$
结束循环
结束过程

3. 性能评估与实验结果

3.1 训练收敛与指标

图3(a)展示了光子脉冲 PPO 算法的训练奖励变化，其中“Raw”表示原始奖励，“MA (100)”表示100轮移动平均。训练早期阶段，奖励迅速上升，并在约1000轮训练后趋于稳定，表明策略学习有效，并逐渐收敛至高质量的路由决策。图3(b)展示了相应的训练损失，该损失起初波动较大，但很快收敛至较低水平，表明优化过程稳定且表现良好。图3(c)–(f)对比了流量负载不断增加时光子脉冲 PPO 与经典 Dijkstra 算法的网络性能。随着基准速率从40兆比特每秒升至200兆比特每秒，光子脉冲 PPO 始终实现更高的吞吐量，达到近6000兆比特每秒，而 Dijkstra 算法的增长则较为有限。同时，光子脉冲 PPO 在所有流量水平下均保持较高的负载均衡性，表明流量分布更加均匀

分布和改进的拥塞避免。在可靠性和时延方面，光子脉冲 PPO 表现出显著低于 Dijkstra 算法的丢包率，即使在高负载条件下也是如此，并且平均时延始终相对稳定地保持在 20 毫秒以下。相比之下，Dijkstra 算法的时延随流量负载增加而急剧上升，在 200 兆比特每秒时超过 80 毫秒。这些结果表明，光子脉冲 PPO 能够有效优化路由路径，从而提高吞吐量、实现网络负载均衡、减少丢包并确保低时延传输。

总体而言，图 3 中的结果证实，所提出的光子脉冲 PPO 算法在训练过程中实现了稳定收敛，并且在多项关键性能指标上始终优于 Dijkstra 算法，验证了其在智能路由优化方面的有效性。

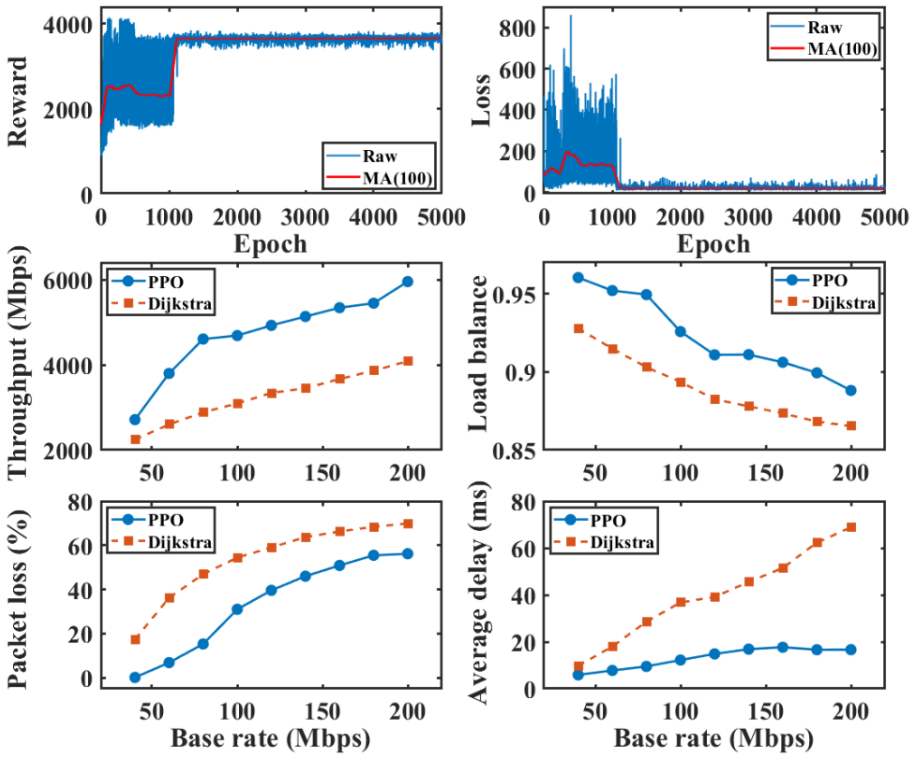

图 3 光子脉冲 PPO 算法的训练奖励和损失曲线。PPO 算法与 Dijkstra 算法的性能比较，包括吞吐量、负

3.2 软硬件协同计算性能

SPGD 优化过程如图 4(a) 所示。为确保与 PSNN 芯片兼容，导出的权重经过裁剪，移除了负权重和偏置项。将隐藏层线性权重映射到 MZI 网格芯片上的相应硬件训练结果如图 4(b) 所示。通过计算训练后的数字权重 A 与硬件映射权重 B 之间的相似度，评估权重映射保真度，

使用如下定义的相似度度量：

$$a = flatten(\mathrm{A}), b = flatten(\mathrm{B}), cf = \frac{a \cdot b}{\|a\|\|b\|} \tag{17}$$

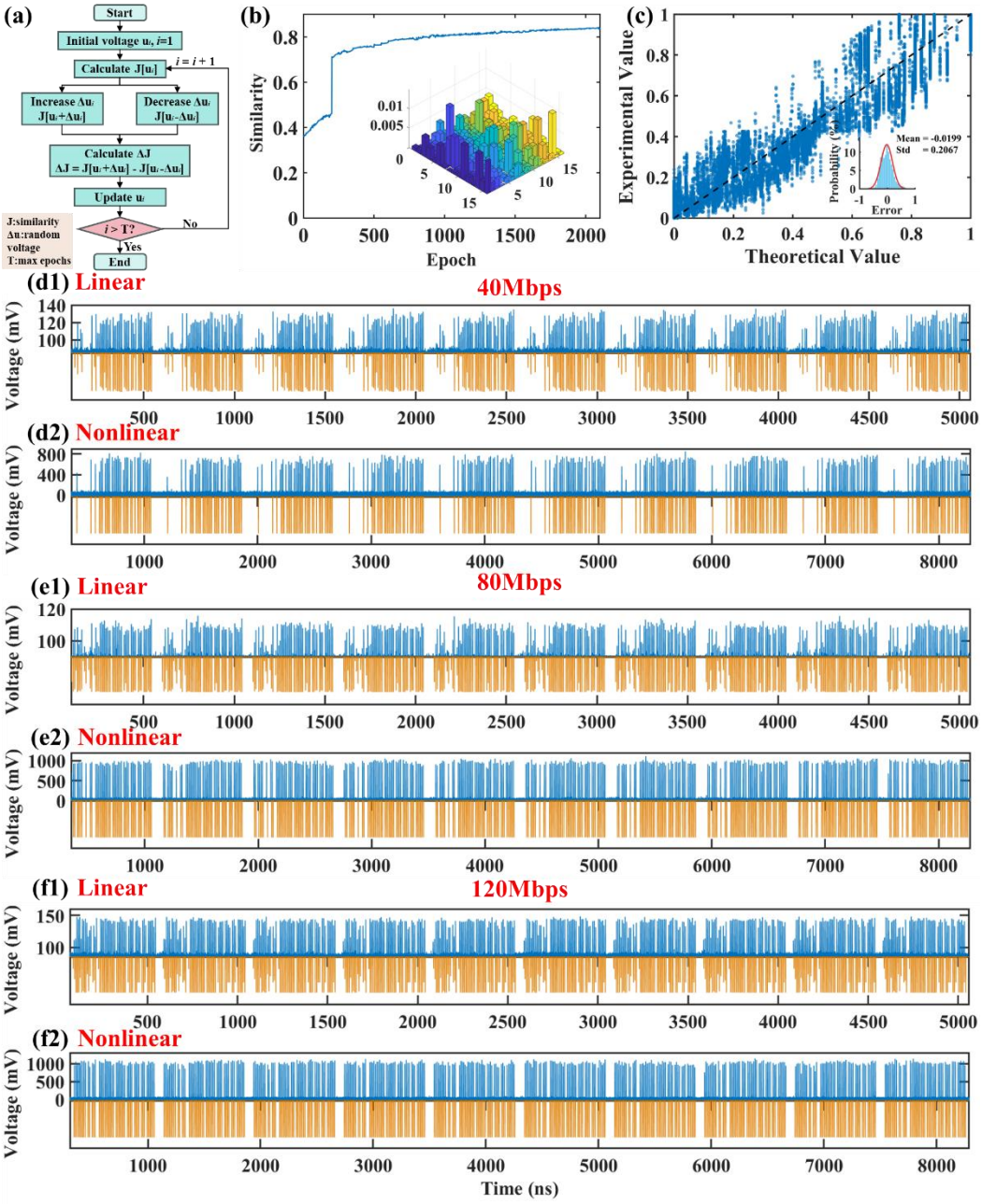

图 4 （a）SPGD算法的优化过程；（b）MZI训练曲线；（c）归一化的理论结果和实验结果；基本速率为

结果表明，在经过约 800 个训练轮次后，相似度接近 0.80，并在 2100 个训练轮次后最终达到 0.84。图 4 (c) 展示了使用 1920 个状态-动作对时，16 个通道（时间步 T=1）的归一化理论结果和实验结果，以及误差分布，所对应的总传输数据量为

30720。误差概率分布曲线表明，误差主要集中在 0 附近，且呈现出符合正态分布的特征。实验结果与理论预测高度一致，表明系统在多通道数据传输过程中保持了出色的稳定性和准确性。图 5 展示了硬件感知软件微调阶段的重新训练结果。微调收敛速度与基线软件训练相当（原始奖励迅速上升，并在约 1000 个训练轮次后趋于稳定；损失快速下降并保持在较低水平），证明了光子脉冲 PPO 的稳定性。尽管奖励略有下降，重新训练后的光子脉冲 PPO 在关键指标（吞吐量、负载均衡、丢包率、平均时延）方面仍优于经典 Dijkstra 算法，证实了其在路由优化中的有效性。

图4(d1)-(d2)、4(e1)-(e2)和4(f1)-(f2)分别对应基本流量速率为40、80和120兆比特每秒时通道10、通道3和通道8的实验结果。对于这三种流量速率，基于MZI的线性计算结果均与算法计算的目标值高度吻合，表明在不同通道和负载条件下，光学矩阵向量乘法具有很高的保真度。对于40兆比特每秒的情况（通道10），如图4(d2)所示，DFB-SA阵列的非线性激活输出在2039.7 ns、2851.89 ns、4463.99 ns和5264.08 ns处出现了四次过激活事件。相应的逐层错误率为0.5371%，最终推理准确率达到99.38%；对于80兆比特每秒的情况（通道3），如图4(e2)所示，仅在4376.49 ns处出现一次过激活事件，因此逐层错误率降至0.4785%，最终推理准确率达到100%；对于120兆比特每秒的情况（通道8），图4(f2)中DFB-SA阵列的非线性激活输出与目标激活高度一致，逐层错误率最低，为0.3516%，最终推理准确率达到100%。我们发现，在这些发生过激活的时刻，注入DFB-SA激光器的信号功率接近激活阈值，因此更容易受到系统中不可避免的扰动和噪声影响。DFB-SA发生过激活的原因是，在目标激活时刻，实际输入光功率略高于实际阈值。

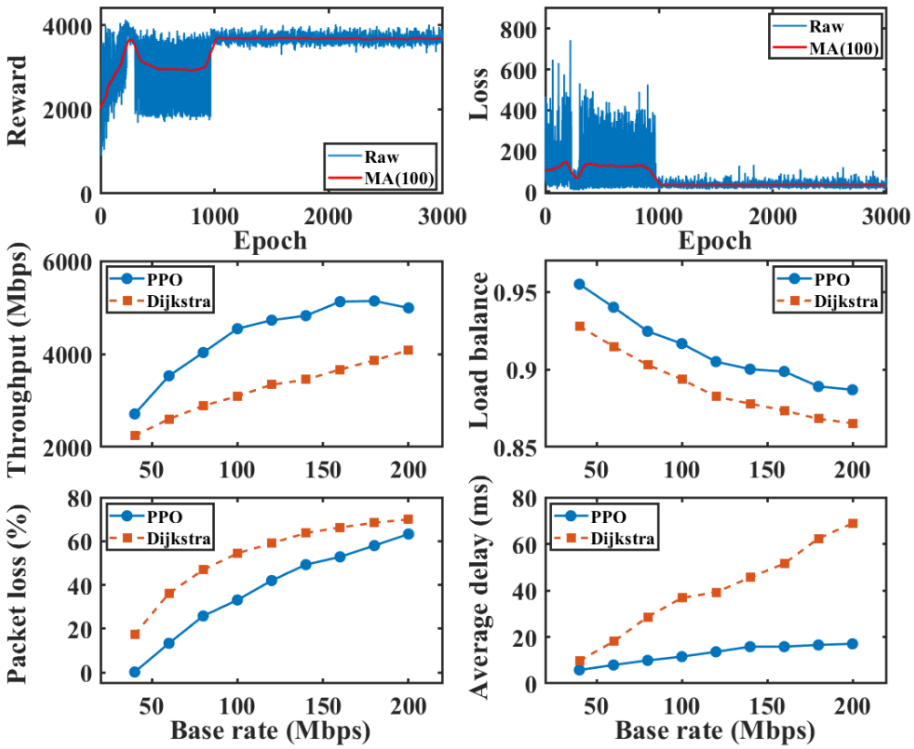

图5. 光子脉冲PPO网络在硬件感知软件微调期间的训练奖励和损失曲线，以及吞吐量、丢包率、平均时延

4. 讨论

本节从两个方面评估所提出的基于光子脉冲强化学习的路由方法的可扩展性：（1）增大脉冲神经网络隐藏层大小（MZI网格规模）；（2）扩大网络拓扑（增大k值的胖树）。本节报告训练动态和路由性能，并将其与经典的Dijkstra算法进行比较。

隐藏层大小（128×128 MZI 网格）。受此前展示的最多 128 个端口的 MZI 网格研究启发 [28]，我们评估了一个 128×128 隐藏层配置。如图 6 所示，原始奖励在训练初期迅速提升，在少量迭代内收敛至约 4000，之后保持稳定，表明该规模下学习效率较高。损失仅在初期出现短暂波动，随后迅速收敛至较低水平，体现出优化过程的稳定性。在所有路由指标上，光子脉冲 PPO 的表现始终优于 Dijkstra 算法。与图 3 中的 16×16 设置相比，128×128 模型收敛更快，而较小的模型由于进行了更充分的早期探索，在后期表现出略好的稳定性。尽管如此，两种配置

收敛后均实现了相近的路由性能，这表明隐藏层大小可以灵活

通过权衡硬件成本、收敛速度和训练稳定性进行选择。

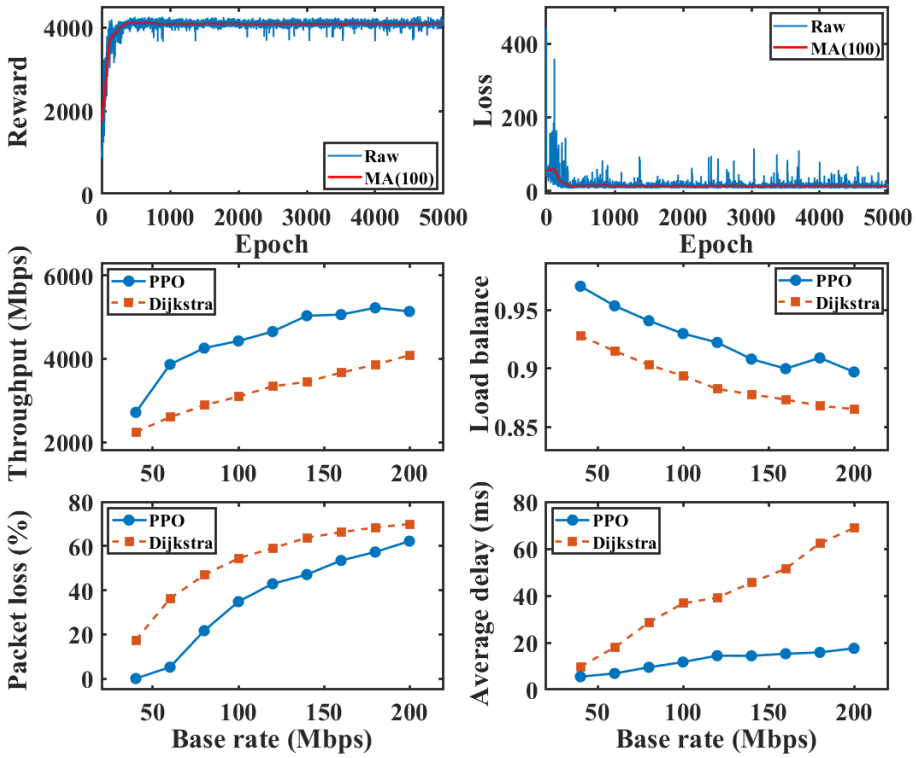

图6 隐藏层规模为 128×128 时光子脉冲 PPO 算法的训练奖励和损失曲线，以及 PPO 算法与 Dijkstra算法

Fat-tree 网络规模（k = 6）。为评估网络层面的可扩展性，我们进一步评估了一个 k = 6 的胖树拓扑，其中包含 54 台主机和 45 台交换机。流量对数量设为 128，每对在训练期间的基准速率为 50 兆比特每秒；评估阶段采用与 k = 4 情况相同的流量速率配置。如图7所示，奖励在约 600 次迭代后收敛，并稳定在 9000 左右，这表明光子脉冲 PPO 能够在更大规模的拓扑中快速学习稳定的路由策略。在性能方面，光子脉冲 PPO 的峰值吞吐量约为 12,000 兆比特每秒，约为 Dijkstra 算法的两倍，这得益于其自适应链路调度能力。在负载均衡方面，两种方法表现相当，而光子脉冲 PPO 始终实现更低的丢包率和平均时延。这些结果表明，随着网络规模扩大，所提出的方法仍保持明显的性能优势和强泛化能力。

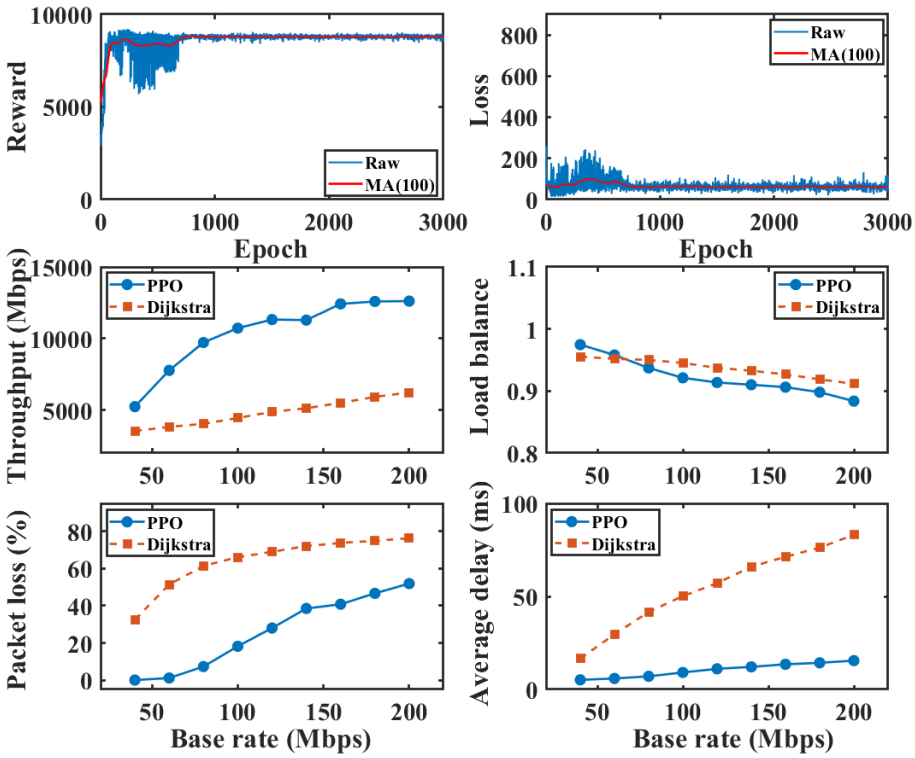

图7 胖树拓扑 k=6 且流量对数量为 128 时，光子脉冲 PPO 算法的训练奖励和损失曲线，以及 PPO 算法与

可扩展性。本方案提出的光子脉冲强化学习芯片可以扩展至更大规模。已有研究表明，硅光 MZI 芯片可扩展至 128 × 128 [28,29]，DFB-SA 激光器阵列可扩展至 150 个通道 [30]。这些进展表明，集成 PSNN 芯片可以进一步扩大，以支持更复杂场景下的智能路由。

5. 结论

我们提出并实现了一种基于光子脉冲 PPO 算法的 SDN 路由优化方法，并针对 SDN 路由开发了专用的光子脉冲 PPO 框架，其中包含明确定义的状态、动作空间和多目标奖励函数。该设计提升了在动态网络环境中的适应能力，而传统路由方法在这类环境中往往难以达到最优。我们在具有代表性的胖树拓扑上验证了该方法。实验结果表明，与经典 Dijkstra 算法相比，光子脉冲 PPO 在吞吐量、丢包率、平均时延和负载均衡方面均取得了显著提升，且在网络负载较高时改善尤为明显，展现出其在复杂场景中的实用价值。此外，我们将脉冲 Actor网络与基于 MZI 的光子线性计算芯片和基于 DFB-SA 的光子脉冲神经元芯片集成，实现了

硬件-软件协同计算，实现低时延、高能效的光子推理。这项工作连接了光子神经形态计算与通信网络，拓展了光子脉冲神经网络的应用范围，并为未来空间–空中–地面一体化网络、大规模数据中心和大规模卫星网络中的路由优化提供了一条前景广阔的超低时延路径。

参考文献

1. Kurose J F, Ross K W. 计算机网络：自顶向下方法，第6版. Pearson, Boston, MA, USA (2012). 2. Guerin R A, Orda A, Williams D. QoS路由机制与OSPF扩展. Proc. IEEE 全球电信会议 **3**, 1903–1908 (1997). 3. Mnih V 等. 通过深度强化学习实现人类水平控制.《自然》 **518**, 529–533 (2015). 4. Lin S C, Akyildiz I F, Wang P, Luo M. 多层分层软件定义网络中的QoS感知自适应路由：一种强化学习方法. Proc. IEEE 服务计算国际会议 25–33 (2016). 5. Valadarsky A, Schapira M, Shahaf D, Tamar A. 学习路由. Proc. ACM HotNets-XVI, 185–191 (2017). 6. Stampa G, Arias M, Sanchez-Charles D, Muntes-Mulero V, Cabellos A. 一种面向软件定义网络路由优化的深度强化学习方法. arXiv:1709.07080 (2017). 7. Chen Y R, Rezapour A, Tzeng W G, Tsai S C. RL-Routing：一种基于深度强化学习的SDN 路由算法. IEEE《网络科学与工程汇刊》 **7**, 3185–3199 (2020). 8. Casas-Velasco D M, Rendon O M C, da Fonseca N L S 等. DRSIR：一种用于软件定义网络路由的深度强化学习方法. IEEE《网络与服务管理汇刊》 **19**, 4807–4820 (2021). 9. Dai B, Cao Y, Wu Z, Dai Z, Yao R, Xu Y. 路由优化遇见机器智能：面向未来网络的展望"

Neurocomputing **459**, 44–58 (2021). 10. Almasan P 等. 深度强化学习与图神经网络相结合：探索路由优化应用案例。Comput. Commun. **196**, 184–194 (2022). 11. Pei X, Sun P, Hu Y, Li D, Chen B, Tian L. 利用深度强化学习实现软件定义网络流量工程的高效路由。Comput. Netw. **241**, 110220 (2024). 12. Wu J, Zhu Z. 基于近端策略优化和 GNN 的软件定义网络智能路由优化。J. Netw. Comput. Appl. **242**, 104249 (2025). 13. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. 近端策略优化算法。arXiv:1707.06347 (2017). 14. Roy K, Jaiswal A, Panda P. 借助神经形态计算迈向基于脉冲的机器智能。《自然》 **575**, 607–617 (2019). 15. Li G, Deng L, Tang H, Pan G, Tian Y, Roy K, Maass W. 脑启发计算：系统综述与未来趋势。Proc. IEEE **112**, 544–584 (2024). 16. Shastri B J 等. 面向人工智能与神经形态计算的光子学。Nat. Photon. **15**, 102–114 (2021). 17. Farmakidis N, Dong B, Bhaskaran H. 集成光子神经形态计算：机遇与挑战。Nat. Rev. Electr. Eng. **1**, 358–373 (2024). 18. Feldmann J, Youngblood N, Wright C D, Bhaskaran H, Pernice W H P. 具有自学习能力的全光脉冲神经突触网络。《自然》 **569**, 208–214 (2019). 19. Xiang S Y, Shi Y, Guo X, Zhang Y, Wang H 等. 基于集成式法布里-珀罗激光器与可饱和吸收体的光子脉冲神经元芯片，实现硬件-算法协同计算。Optica **10**, 162–171 (2023). 20. Shi Y C, Xiang S Y, Guo X X, Zhang Y H, Wang H 等. 基于自脉动DFB与可饱和吸收体的光子集成脉冲神经元芯片。Photon. Res. **11**, 1382–1389 (2023). 21. Xiang S Y, Shi Y C, Guo X X, Zhang Y H 等. 用于卷积脉冲神经网络的光子集成神经突触核

22. Yang Z, Zhang T, Dai J, Xu K. 用于路径规划中强化学习的基于可调偏置的光学神经网络。Opt. Express **32**, 18099–18112 (2024).23. Li X K, Ma J X, Li X Y, Hu J J, Ding C Y 等。采用混合架构光子集成电路实现高效强化学习。Nat. Commun. **15**, 1044 (2024).24. Kreutz D, Ramos F M V, Veríssimo P E, Rothenberg C E, Azodolmolky S, Uhlig S. 软件定义网络：全面综述。Proc. IEEE **103**, 14–76 (2015).25. Masoudi R, Ghaffari A. 软件定义网络：综述。J. Netw. Comput. Appl. **67**, 1–25 (2016).26. Xiang S Y, Chen Y H, Zhao H W 等。用于脉冲强化学习的非线性光子神经形态芯片。arXiv:2508.06962 (2025). 27. Wan Y, Liu X, Wu G, Yang M, Yan G, Zhang Y, Wang J. 用于片上光处理器的高效随机并行梯度下降训练。Opto-Electron. Adv. **7**, 230182 (2024).28. Ahmed S R, Baghdadi R, Bernadskiy M 等。通用光子人工智能加速。《自然》**640**, 368–374 (202

5).

29. Hua S、Divita E、Yu S 等。具有超低延迟的集成式大规模光子加速器。《自然》**640**，361–367（2025）。30. Zhang Y 、 Sun Z 、 Xiao R 、 Chen X。基于非对称多量子阱的 150 通道 REC-DFB 激光器阵列的实验演示。IEEE J. Quantum Electron. **61**，2200307（2025）。

致谢

我们衷心感谢国家自然科学基金（编号：62535015、62575231）、中央高校基本科研业务费（QTZX23041）以及西安电子科技大学交叉探索专项资助项目（TZJH2024009）提供的经费支持。

作者贡献

S Y Xiang：概念构思、仿真、经费获取、修订；Y H Chen：实验、可视化、测量、数据整理、撰写初稿；L Zheng：方法学、验证、数据整理；Z C Tu、X T Zeng、M T Yu、S Wang：实验、测量、可视化；Y H Zhang、X X Guo：调查研究、数据整理、修订；W T Pan、Y Hao：监督。

竞争利益

作者声明不存在竞争性财务利益。