An Ultralow-latency Constrained Quadratic Program (QP) Solver based on an Integrated Photonic Recurrent Neural Network

Yusuf O. Jimoh,[1,*] Jiawei Zhang,[1] Hyuma Umeda,[1] Bhavin J. Shastri,[2] and Paul R. Prucnal[1]

[1] *Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544, USA*
[2] *Department of Physics, Engineering Physics & Astronomy, Queen's University, Kingston, Ontario K7L 3N6, Canada*

[*] *yjimoh@princeton.edu*

Abstract: Quadratic program solvers for model predictive control require multiple iterations and are therefore time-consuming. Here, we demonstrate a photonic neural network performing predictive control in under 400 nanoseconds, a thousand times improvement over today's computers. © 2025 The Author(s)

Quadratic program (QP) optimization is a standard tool across many fields, including portfolio optimization in finance, support vector machines in machine learning, and model predictive control in robotics. In latency-critical applications, particularly real-time control, it becomes increasingly crucial to solve QP instances within milliseconds, microseconds, or even nanoseconds, depending on the application. However, applications requiring the latter two timescales are largely beyond the reach of even the most powerful digital electronic solvers. First, the most common QP solvers that run on digital hardware utilize iterative, second-order algorithms such as the interior-point method (IPM) or the active-set method, which are limited by their computational time complexity ($O(n^2) - O(n^3)$ as the problem size, n, increases). Although various mathematical strategies have been proposed to reduce the computational complexity per iteration, such as the fast gradient-descent method [1], these approaches often come with significantly slower convergence. Alternatively, other studies focus on improving the efficiency and throughput of digital computing hardware, including the development of graphics processing unit (GPU) [2]- and field-programmable gate array (FPGA) [3]-based accelerators. While these advanced digital processors can offer massive parallel compute cores and lower power consumption, both are still fundamentally limited by their clock frequencies, and GPUs, particularly, incur high minimum latencies due to memory-access times, rendering them useless in many latency-critical applications.

In this work, we demonstrate a fully integrated photonic QP solver for implementing real-time model predictive control (MPC), an optimal control algorithm widely used in robotics and industrial plant control. MPC is an advanced control strategy that predicts future behavior using a mathematical model and optimizes control actions subject to input and state constraints. As shown in Fig. 1a and 1b, for each sampling time step k, MPC can be formulated as a convex constrained QP (CQP) problem, and the solution of the CQP is very accurately approximated by the steady state of an appropriately configured CTRNN [4]. CTRNN-based solvers can solve CQP problems with very low, near-deterministic latency, which favors their use in real-time systems where the optimization is embedded in a feedback control loop. Integrated photonics, compared to digital and analog electronics, has been demonstrated to be an excellent platform for high-bandwidth, low-latency processing, thereby motivating the development of photonic CTRNNs to achieve the lowest-latency solvers. The details of mapping the CQP problem to a photonic CTRNN are presented in our previous work [4].

The experimental setup is illustrated in Fig. 1c–f. The photonic integrated circuit (PIC) implements a two-neuron continuous-time recurrent neural network (CTRNN) with a footprint of less than 0.95×0.9 mm^2. Each neuron is defined by three synaptic weights: two recurrent connections and one external input. The linear front-end and nonlinear backend of the microring resonator (MRR)-based wavelength-division-multiplexed (WDM) photonic neurons are extensively detailed in [5]. Both neuron outputs are coupled to a common optical bus waveguide, which is then split equally into three channels and fed back to the neurons, with the last channel reserved for real-time state monitoring. The nonlinear backend optical-electrical-optical (OEO) neurons exhibit a measured bandwidth of about 10 MHz, corresponding to a neuron time constant of 0.1 μs. In this experiment, we use one neuron as the solver neuron and the other for minimum-constraint enforcement.

We experimentally demonstrate MPC on a proof-of-concept one-dimensional (1D) reference-tracking task. As shown in Fig. 2a, the tracking task is to control a virtual plant to follow a 1D reference trajectory (blue, solid curve) while respecting the minimum position constraint of –0.7 (black, dashed curve). When the tracking task requires a sufficiently high sampling rate (e.g., 1 MHz), a conventional CPU-based QP solver (yellow, dotted curve) based on

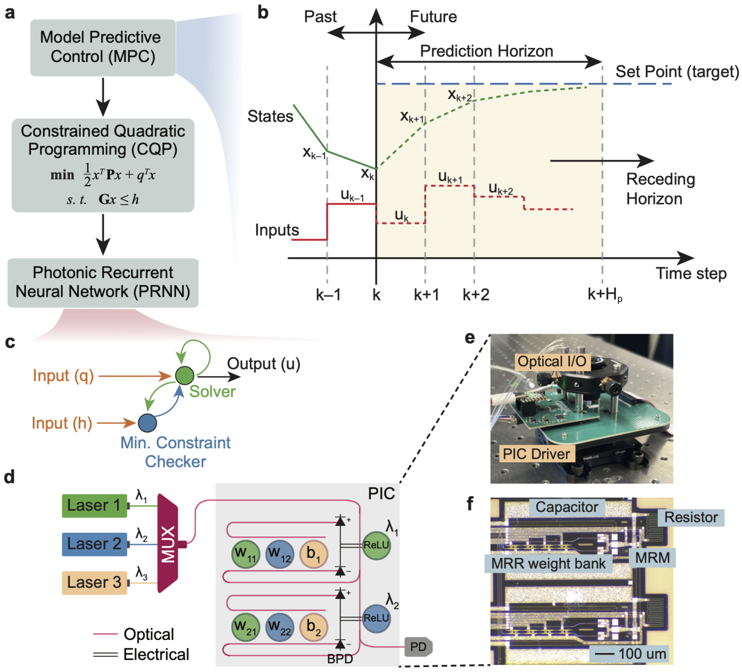

Fig. 1. Implementation and experimental setup of a photonic QP solver. **a** Mapping from MPC to QP problem to a photonic CTRNN. **b** Concept of MPC. **c** Configuration of the CTRNN. **d** Schematic of the two-neuron experiment. **e** Photo of the experimental setup. **f** Micrograph of the PIC.

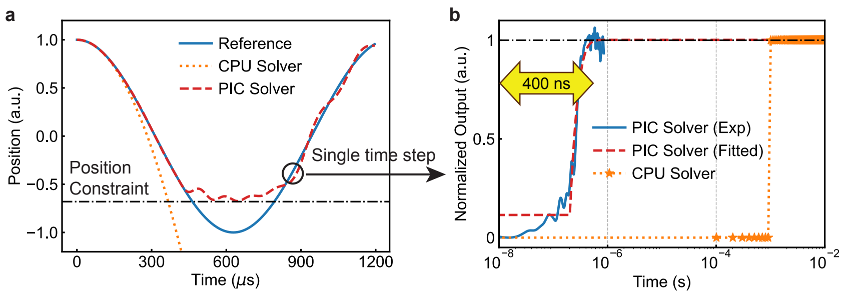

Fig. 2. Experimental results of photonic QP solver for model predictive control. **a** 1D constrained position tracking of a reference trajectory. **b** Comparison of convergence time of a PIC and CPU solver for a single sampling time step.

IPM results in significant deviation from the reference trajectory. In contrast, our photonic QP solver (red, dashed curve) accurately tracks the reference while satisfying the specified constraint. As shown in Fig. 2b, the output of our photonic QP solver converges to its steady state within 400 ns, achieving more than three orders of magnitude latency improvement compared to the CPU solver (about 1 ms).

In conclusion, we demonstrate the first fully integrated photonic QP solver and achieve a convergence latency of 400 ns, the best to date among QP solvers reported in the literature. We estimate the convergence time to scale as $O(\sqrt{n})$ as the problem dimension, n, increases. Also, the minimum latency can be further reduced to 10s of nanoseconds by optimizing the bandwidth of OEO neurons. Our work fundamentally addresses the challenges faced by current QP solvers in both latency and computational complexity, representing a significant milestone for PICs in enabling critical, real-world applications.

References

1. Peyrl, H., *et al.*, *Control Eng. Practice* **33**, 22–34 (2014)
2. Schubiger, M., *et al.*, *J. Parallel Distrib. Comput.* **144**, 55–67 (2020)
3. Lucia, S., *et al.*, *IEEE Trans. Ind. Inform.* **14**, 137–145 (2017)
4. De Lima, T., *et al.*, *J. Lightwave Technol.* **37**, 1515–1534 (2019)
5. Tait, A., *et al.*, *JSTQE* **22**, 312–325 (2016)