Advancing New Hardware Machine Learning Based on Reservoir Computing

Daniel Soh, Associate Professor
College of Optical Sciences, the University of Arizona
ECE Seminar February 2025

THE UNIVERSITY OF ARIZONA
Wyant College of Optical Sciences

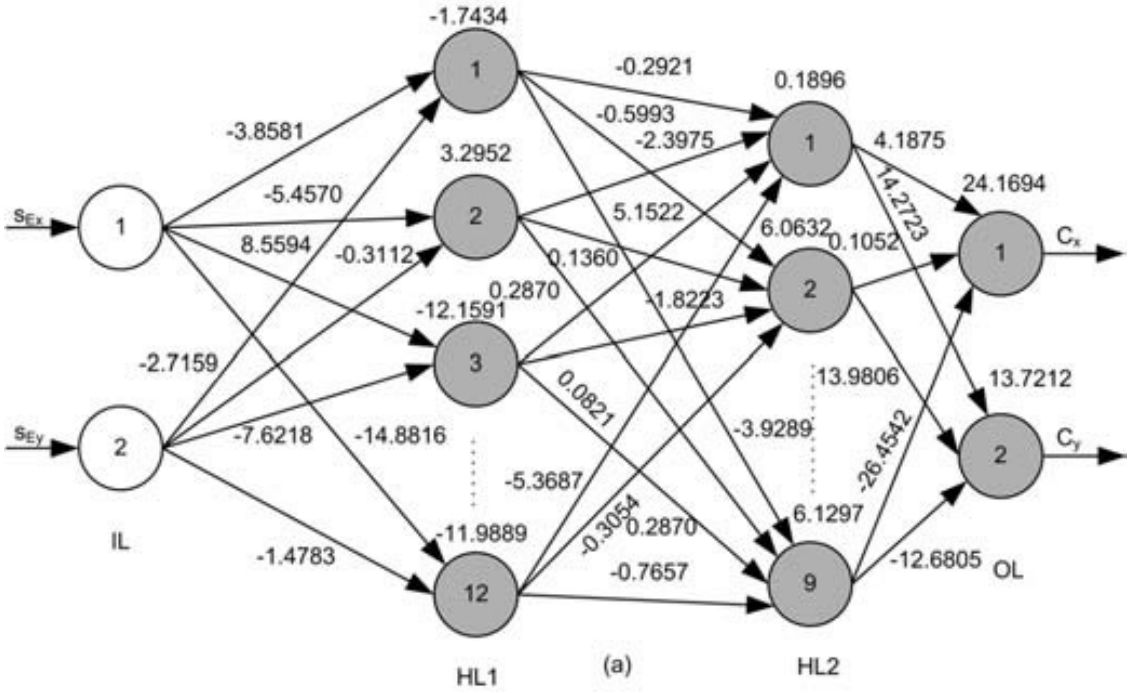
Contents

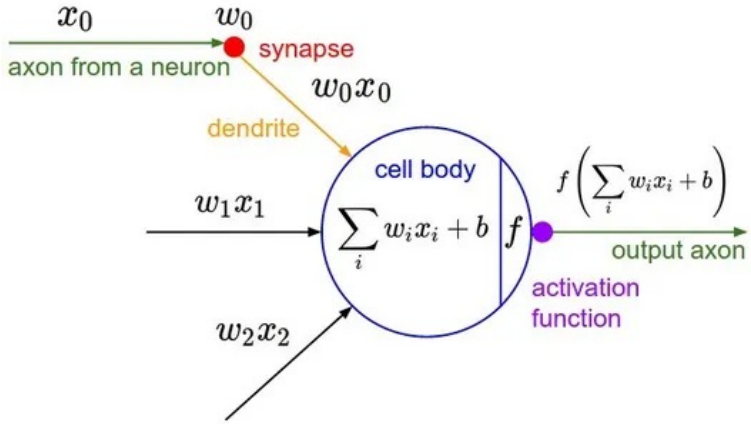
- New Paradigm 1 – Reservoir Computing
- New Paradigm 2 – Hardware Machine Learning
- New Paradigm 3 – Quantum Hardware Reservoir Computing
- Theoretical Advancements We Contributed
- Example Results of Optical and Quantum Hardware Reservoir Computing

THE UNIVERSITY OF ARIZONA
Wyant College of Optical Sciences

Neural Network Machine Learning

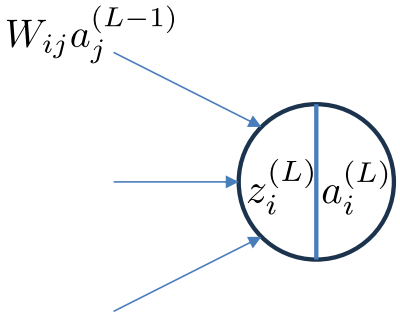
- Combination of linear and nonlinear connection of neurons.





Neural Network Machine Learning

- Supervised learning - back propagation weight training to reduce the cost function C_0



Cost function C_0 is a function of $a^{(L)}$, which is a function of $z^{(L)}$ …

How much does a nudge to $w^{(L)}$ change $z^{(L)}$?

How much does that nudge to $z^{(L)}$ change $a^{(L)}$?

$$\frac{\partial C_0}{\partial w^{(L)}} = \frac{\partial z^{(L)}}{\partial w^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial C_0}{\partial a^{(L)}}$$

How much does *that* nudge to $a^{(L)}$ change C_0?

- Weight training update rule

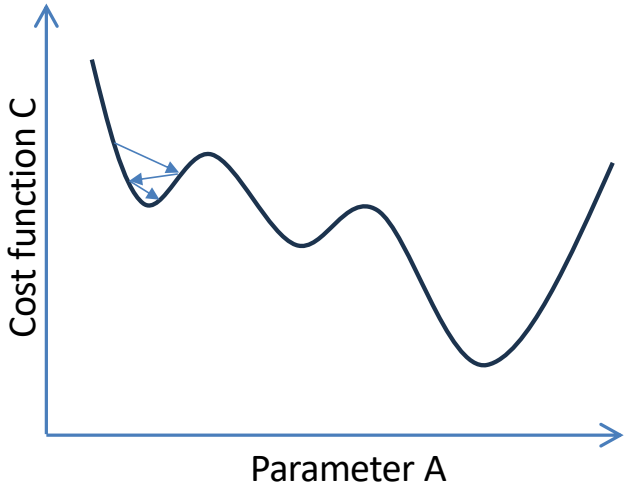
$$w'^{(L)} = w^{(L)} - \alpha \frac{\partial C_0}{\partial w^{(L)}}$$

THE UNIVERSITY OF ARIZONA
Wyant College of Optical Sciences

Numerical Instability in Deep Neural Networks

- **Numerical Instability in Backpropagation**
  - **Vanishing & Exploding Gradients**
    - Deep networks can suffer from extremely small or large gradients, slowing or destabilizing training.
  - **Repeated Multiplications**
    - Long chains of matrix multiplications accumulate numerical errors, causing instability.
  - **Some Mitigation Strategies Developed**
    - Weight initialization (e.g., Xavier, He), normalization layers (e.g., BatchNorm), and architectural innovations (e.g., residual connections).
- **Difficulty in Finding Global Minima**
  - **High-Dimensional Loss Landscape**
    - Millions of parameters create a complex error surface with many local minima, saddle points, and plateaus.
  - **Gradient Descent Challenges**
    - Vanilla gradient descent may converge slowly or get stuck; advanced optimizers (Adam, RMSProp) help navigate complex landscapes.
  - **Practical Observations**
    - Despite the complexity, suitable hyperparameters and architectures often yield good "enough" solutions.

[He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition*. In Proceedings of the IEEE CVPR.]



THE UNIVERSITY OF ARIZONA
Wyant College of Optical Sciences

State-of-the-art AI today: ChatGPT

- ChatGPT-3 (ChatGPT-4 hardware details are kept secret)
  - Number of weight parameters – 175 billion parameters
  - Number of layers – 96 layers
  - Number of chips used – several thousands of Nvidia V100 GPUs. Assuming each GPU has 640 Tensor Cores, the total number of individual processing cores is in the range of 10^6 cores.
  - **Required energy for training – 1287 MWh, equivalent to the energy consumed by 120 American households in a year.**
  - Required energy for a single query – a few mWh per query
  - Data center hosting a CHATGPT processor – 691 square feet computing cores + additional infrastructure for cooling, networking, aisles for access (several thousands of square feet altogether).

[Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165.]

- Another bottleneck of CHATGPT and other AIs
  - Excessive consumption of an astronomical amount of training data without a stellar performance.

THE UNIVERSITY OF ARIZONA
Wyant College of Optical Sciences

State-of-the-art AI today: Deepseek R1

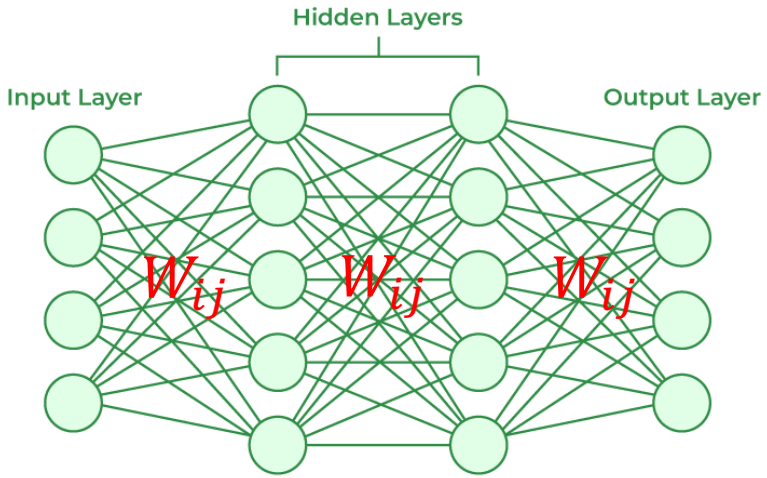
- Adaptive, Multi-Modal Architecture & Knowledge Distillation
  - Dynamically adjusts to data patterns, fusing text, image, and tabular inputs.
  - Leverages teacher–student paradigms to deliver efficient, high-performing models.
- Automated, Real-Time Pipeline
  - Streamlines data preprocessing, training, and deployment for rapid iteration.
  - Optimized for low-latency predictions in both cloud and edge environments.
- Transparent & Extensible Framework
  - Built-in Explainable AI (XAI) to illuminate decision-making processes.
  - Modular design enables quick experimentation, expansion, and integration.

- Training cost to accomplish a similar performance of ChatGPT-O1 (claim)
  - $6M (c.f., estimate of ChatGPT 4 training cost >$100M)

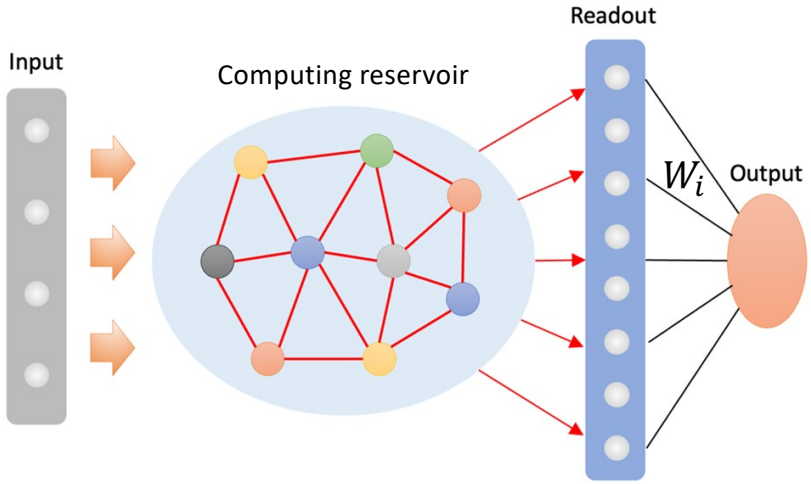
THE UNIVERSITY OF ARIZONA
Wyant College of Optical Sciences

New Paradigm 1 – Reservoir Computing

- Where the training happens



Neural Network: training happens for all weights



Reservoir Computing: Training happens only in the readout layer.

- Enormous reduction in training, relying on the built-in complexity of the computing reservoir
- Such simple training scheme leads to *mathematically provable performance.*

THE UNIVERSITY OF ARIZONA
Wyant College of Optical Sciences

Universality of computing

- **Stone-Weierstrasss Theorem**
  - Continuous functions on a compact set can be uniformly approximated by polynomial (or certain algebraic) functions.
- **Fading Memory**
  - System “forgets” distant past inputs over time (bounded influence of older inputs).
  - Ensures stability and a well-defined “state” for approximation.
- **Polynomial Algebra**
  - Readout or output functions that can approximate polynomials (or continuous functions) on the system’s state space.
  - Connects directly to the Stone–Weierstrass framework: if you can implement polynomials, you can approximate a broad class of continuous functions.
- **Separability (Topological Constraint)**
  - Two different inputs lead to different outputs, allowing for dense approximations.
  - Ensures the space can be “finely approximated” by countable sets of polynomial expansions.

[Lukoševičius, M., & Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. Computer Science Review, 3(3), 127–149.]

THE UNIVERSITY OF ARIZONA
Wyant College of Optical Sciences

Types of Reservoir Computers that has Mathematically Proven Universality

- **Echo State Network**
  - Discrete-time recurrent neural network
  - Echo state property – the internal state asymptotically depend only on inputs, rather than initial conditions.

[Grigoryeva, L., & Ortega, J.-P. (2018). Echo state networks are universal. Neural Networks, 108, 495–508.]

- **Liquid State Machines**
  - Continuous-time, spiking-neuron counterpart to ESNs. Instead of discrete-time updates, it uses biological or biologically inspired spiking dynamics in the “liquid” (the reservoir)Ensures stability and a well-defined “state” for approximation.

[Maass, W., Natschläger, T., & Markram, H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. Neural Computation, 14(11), 2531–2560.]

- **Nonlinear Delay-based Reservoir Computing**
  - A single physical or simulated nonlinear node whose output is time-multiplexed to create a high-dimensional “virtual” state. Examples include optical or electronic systems with a time-delay feedback loop.

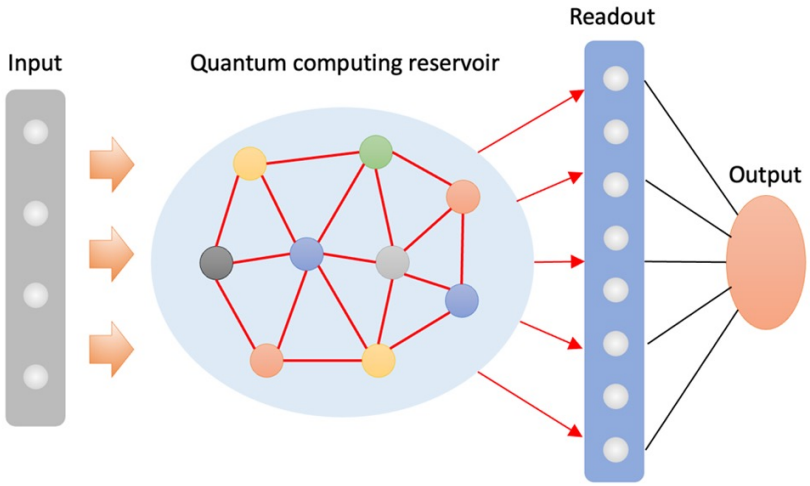
[Grigoryeva, L., & Ortega, J.-P. (2019). Differentiable reservoir computing. Journal of Machine Learning Research, 19, 1–43]

- **Stochastic Reservoir Computing (my group’s recent contribution)**
  - Uses probabilities as node states, providing massive exponential scalability

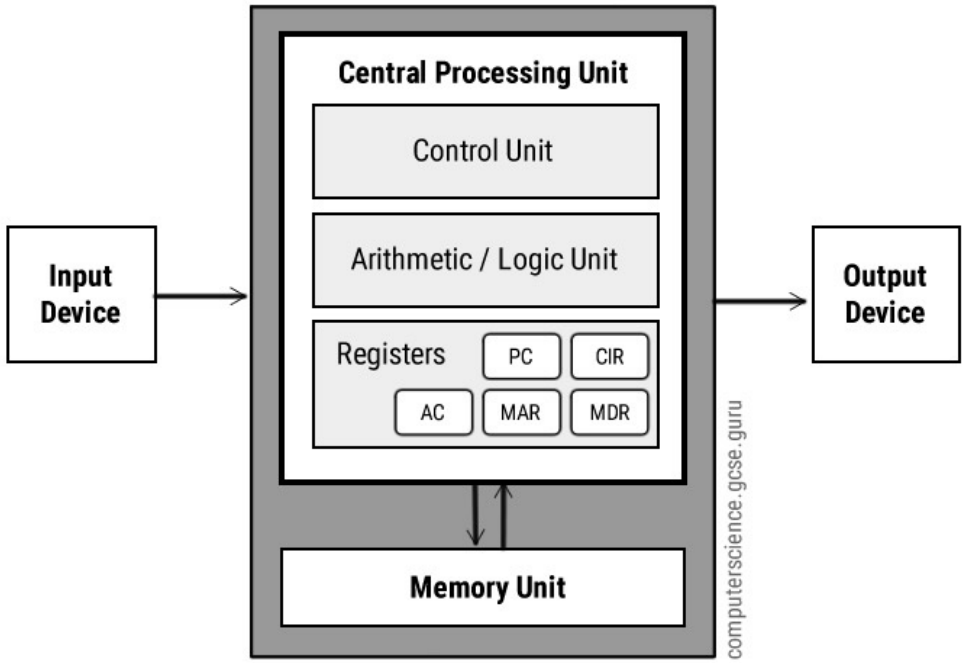
[Ehlers, P. J., Nurdin, H. I., & Soh, D. (2024). Stochastic Reservoir Computers. arXiv preprint arXiv:2405.12382.]

THE UNIVERSITY OF ARIZONA
Wyant College of Optical Sciences

New Paradigm 2 – Hardware Machine Learning



Hardware Machine Learning



Software Machine Learning

THE UNIVERSITY OF ARIZONA
Wyant College of Optical Sciences

Hardware ML vs. Software ML

Aspect	Hardware ML	Software ML
Speed & Throughput	- Specialized acceleration offers high parallelism and fast computations. - Can handle massive workloads efficiently.	- Dependent on general-purpose CPUs or cloud-managed instances. - May experience performance bottlenecks for very large models.
Energy Efficiency	- Custom hardware can achieve high performance per watt, reducing operational costsover time. - Lower power usage at scale.	- Less power-efficient due to overhead on general-purpose hardware. - Potentially higher energy costs for large-scale deployments.
Latency	- Low-latency inference possible with direct hardware implementation. - Beneficial for real-time systems (e.g., autonomous vehicles).	- Network overhead or shared resources in cloud environments may cause variable response times. - Less predictable real-time performance.
Flexibility & Upgradability	- Less adaptable to evolving ML algorithms. - Upgrading means new chip design or FPGA reconfiguration, which can be slow or costly.	- Highly flexible: can quickly update frameworks, retrain, or switch models. - Rapid prototyping and easier iteration.
Use Case Suitability	- Ideal for large-scale or low-latency ML inference (e.g., data centers, edge devices needing real-time response).	- Best for fast development, frequent updates, research, or moderate workloads (especially in the cloud).

THE UNIVERSITY OF ARIZONA
Wyant College of Optical Sciences

All Optical Echo State Network Reservoir Computing

- Echo state network (a type of reservoir computing)
  - Stimulated Brillouin Optical Amplifier activation function

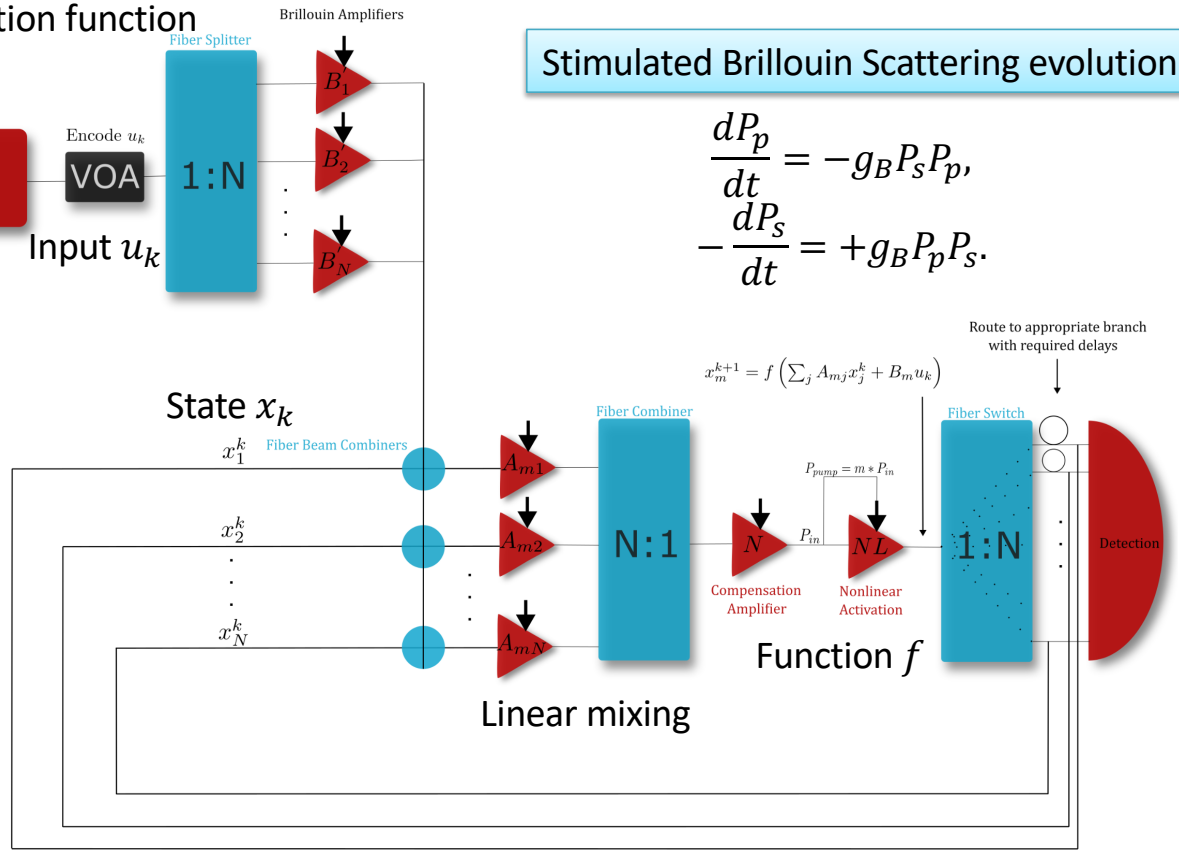
Echo state network evolution

State at t=k+1 — Nonlinear activation — Linear mixing — Input coupling at t=k

$$x_{k+1} = f(Ax_k + Bu_k)$$

$$y_k = W^T x_k + C$$

Output at t=k — Weight to be trained



Stimulated Brillouin Scattering evolution

$$\frac{dP_p}{dt} = -g_B P_s P_p,$$

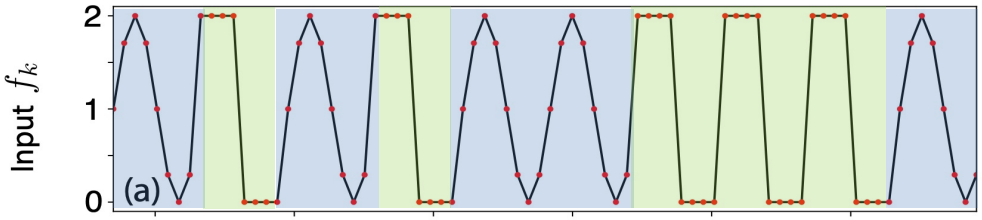
$$-\frac{dP_s}{dt} = +g_B P_p P_s.$$

[Ishwar Kaushik, Peter Ehlers, and Daniel Soh, "All Optical Echo State Network Reservoir Computing," manuscript in preparation]

THE UNIVERSITY OF ARIZONA
Wyant College of Optical Sciences

Training and testing the reservoir computer

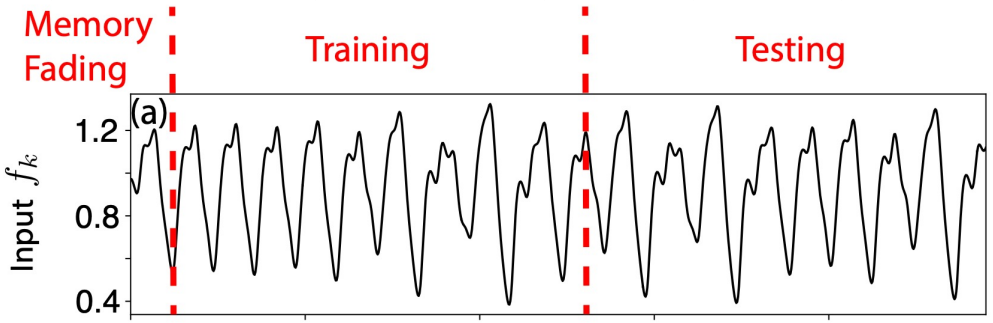
- Tasks and testing properties
  - Sine-Square Input Classification: degree of nonlinearity in the reservoir (c.f. Polynomial algebra)



  - Mackey-Glass Chaotic System Behavior Prediction: dynamic memory, nonlinearity, and separability

$$\frac{dP(t)}{dt} = \frac{\beta_0 \theta^n P(t-\tau)}{\theta^n + P(t-\tau)^n} - \gamma P(t)$$

[Mackey, M. C., & Glass, L. (1977). Oscillation and chaos in physiological control systems. Science, 197(4300), 287-289.]

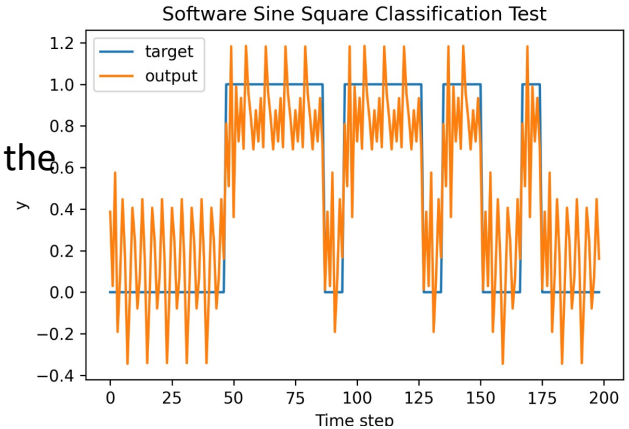


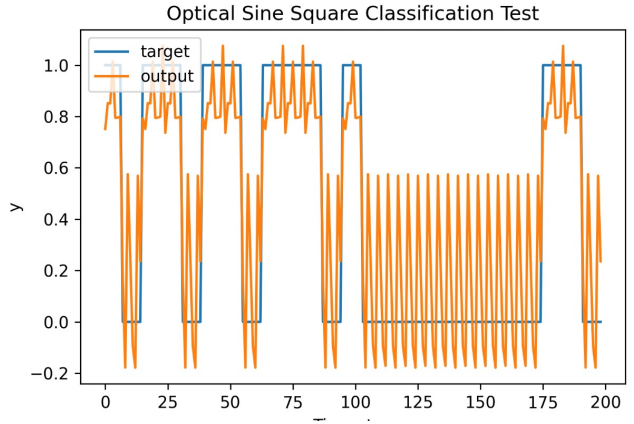
- Training and testing sessions
  - The first part is used to perform supervised training on the machine.
  - The second part is used to measure the performance from the error.

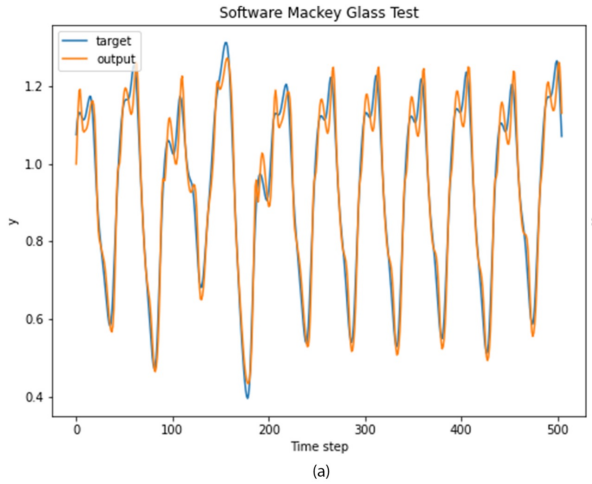
THE UNIVERSITY OF ARIZONA
Wyant College of Optical Sciences

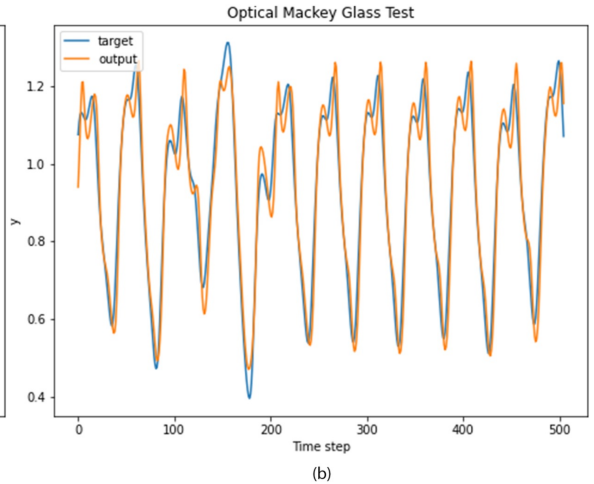
All Optical Echo State Network Reservoir Computing

- Training tests
  - Sine-Square classification test: testing nonlinearity capability of the reservoir
  - Mackey Glass Chaotic Dynamics simulation test: testing the capability to track highly chaotic time-series behavior







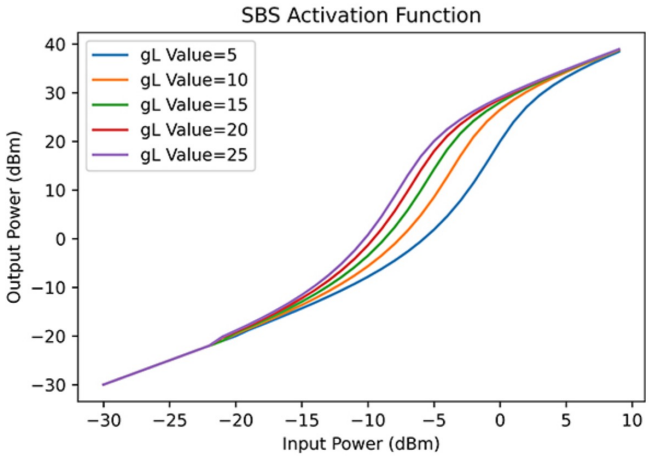


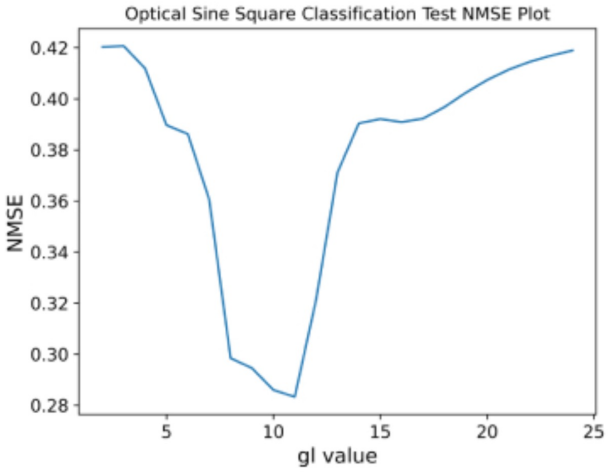
[Ishwar Kaushik, Peter Ehlers, and Daniel Soh, "All Optical Echo State Network Reservoir Computing," manuscript in preparation]

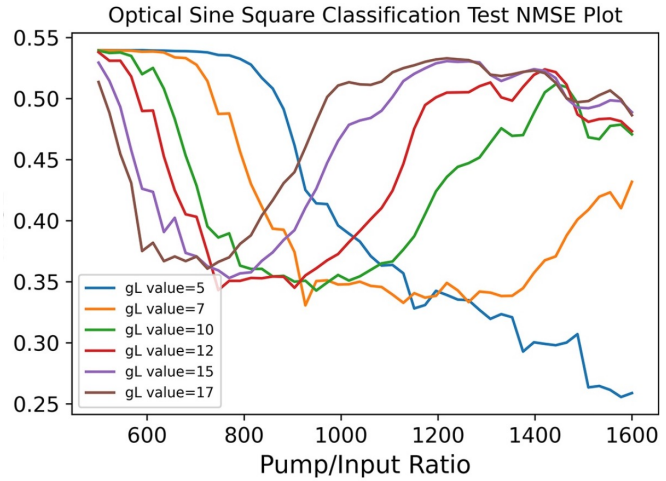
THE UNIVERSITY OF ARIZONA
Wyant College of Optical Sciences

All Optical Echo State Network Reservoir Computing

- Reservoir hardware properties affect the performance greatly.
  - Brillouin nonlinear gain parameter (gL value)
  - Input power level to the Brillouin amplifier







(NMSE: Normalized Mean Square Error)

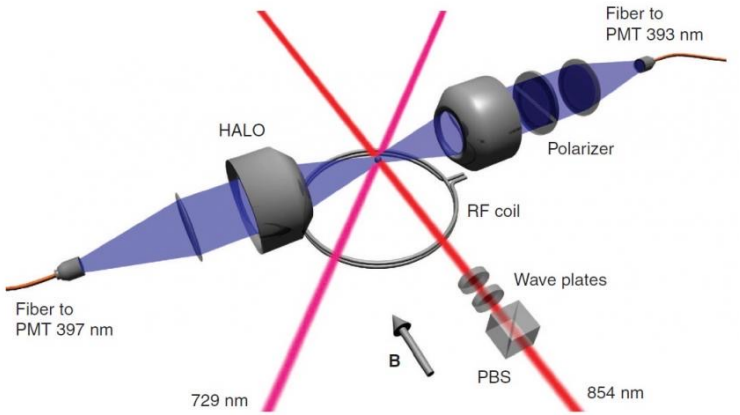
[Ishwar Kaushik, Peter Ehlers, and Daniel Soh, "All Optical Echo State Network Reservoir Computing," manuscript in preparation]

New Paradigm 3 – Quantum Reservoir Computing

- Minimalistic Hardware for Energy Efficiency
  - Reduced system size lowers the energy cost in both training and operation.
  - Suitable for **distributed edge-computing** with drastically reduced network communication
- Exponential Hilbert Space = Large Reservoir
  - A small quantum system harnesses an **exponentially growing state space**.
  - Enormous effective capacity enables powerful computation with fewer physical layers.
- Decoherence Enables Fading Memory
  - Conventional quantum computing views decoherence as detrimental.
  - In reservoir computing, **partial decoherence is essential**, providing the “forgetting” or fading memory necessary for processing temporal data.

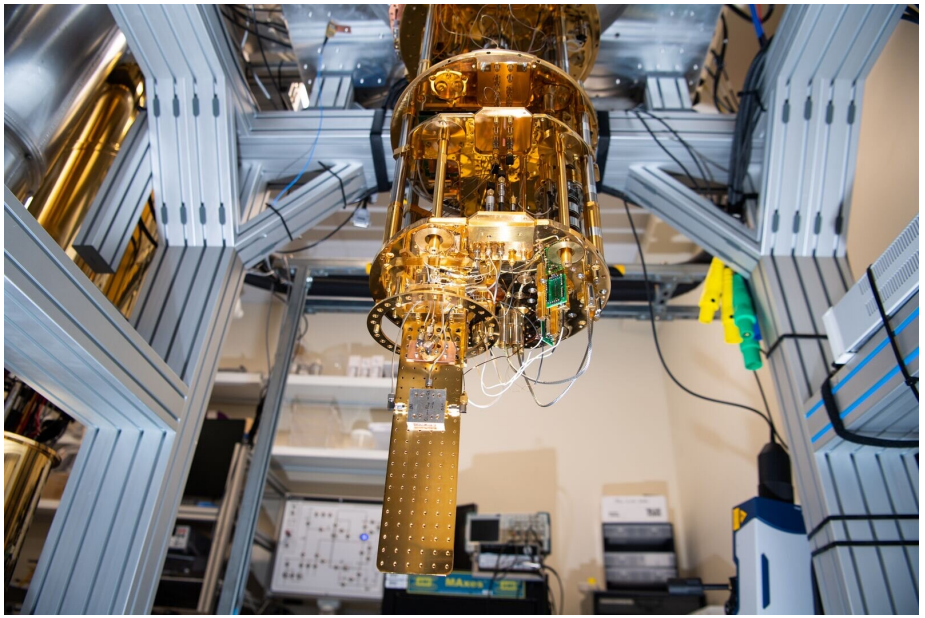
THE UNIVERSITY OF ARIZONA
Wyant College of Optical Sciences

Queries and gate operations in QC are hard



Conversion of classical input to qubits: quantum state preparation

Q: is it fair to count queries and gate operations, and then compare with classical computers on an equal footing?



- Gate operations need to prevent decoherence as much as possible.
- Error correction overhead is **ENORMOUS:** typical ratio of physical vs. logical qubits ~ 1000:1
- Readout is difficult.

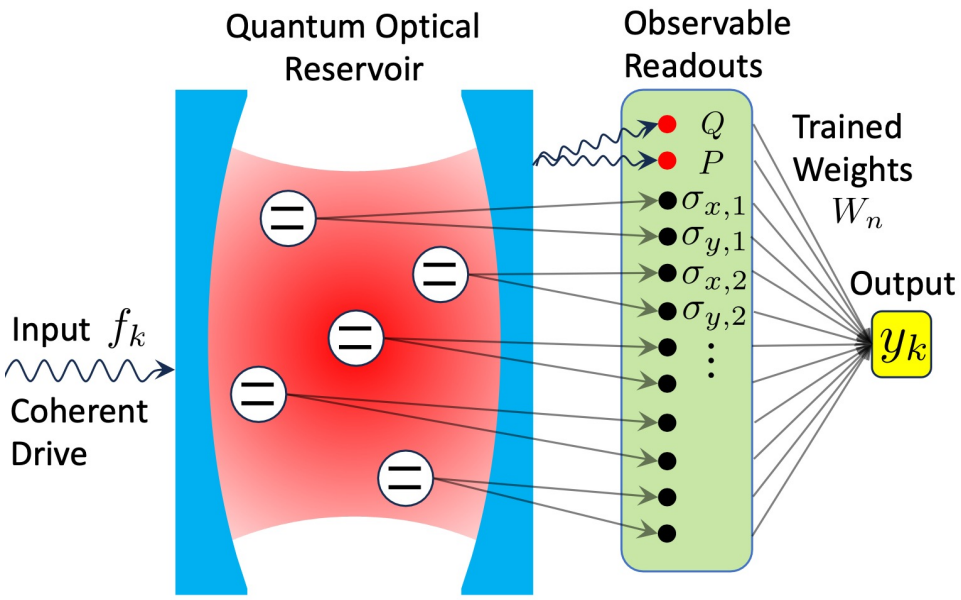
THE UNIVERSITY OF ARIZONA
Wyant College of Optical Sciences

Reservoir Computing using Noisy Intermediate-scale Quantum (NISQ)

- **Harnessing "Useful" Noise**
  - Partial Decoherence is Beneficial: In quantum reservoir computing, slight decoherence is essential for the fading memory effect.
  - **No Need for Full Error Correction**: The high overhead of fault-tolerant quantum computing can be avoided; noise helps rather than hinders.
- **Intermediate Scale = Sufficiently Large Hilbert Space**
  - Exponential Scaling: Even a modest number of qubits provides a vast state space.
  - Practical Complexity: **NISQ devices strike a balance** between being large enough to exhibit rich dynamics yet still within current technological reach.
- **Accessible, Near-Term Quantum Technology**
  - **Immediate Experimental Realization**: Existing NISQ setups (e.g., superconducting qubits, trapped ions) can serve as quantum reservoirs.
  - Pathway to Scalable Approaches: Insights gained from NISQ platforms inform next-generation hardware without waiting for fully fault-tolerant machines.

THE UNIVERSITY OF ARIZONA
Wyant College of Optical Sciences

Few-atom Reservoir Computing



(Continuous measurement with measurement back-action)

Hamiltonian:

$$H_0 = \omega_c c^\dagger c + \sum_i \omega_i \sigma_i^\dagger \sigma_i + \sum_i g_i \left(c^\dagger \sigma_i + c \sigma_i^\dagger \right),$$

Field — Atoms — Coupling

$$H_1(t) = i\epsilon f(t) \left(c - c^\dagger \right),$$

Input field coupling

Quantum evolution:

Collapse operators (quantum back action)

$$\frac{d\rho}{dt} = -i\left[H_0 + H_1(t), \rho\right] + 2\mathcal{D}\left[\sqrt{\kappa_c} c\right]\rho + 2\sum_i \mathcal{D}\left[\sqrt{\kappa_i}\sigma_i\right]\rho,$$

$$\mathcal{D}[a]\rho = a\rho a^\dagger - \frac{1}{2}\left(a^\dagger a \rho + \rho a^\dagger a\right)$$

Measurement (observables):

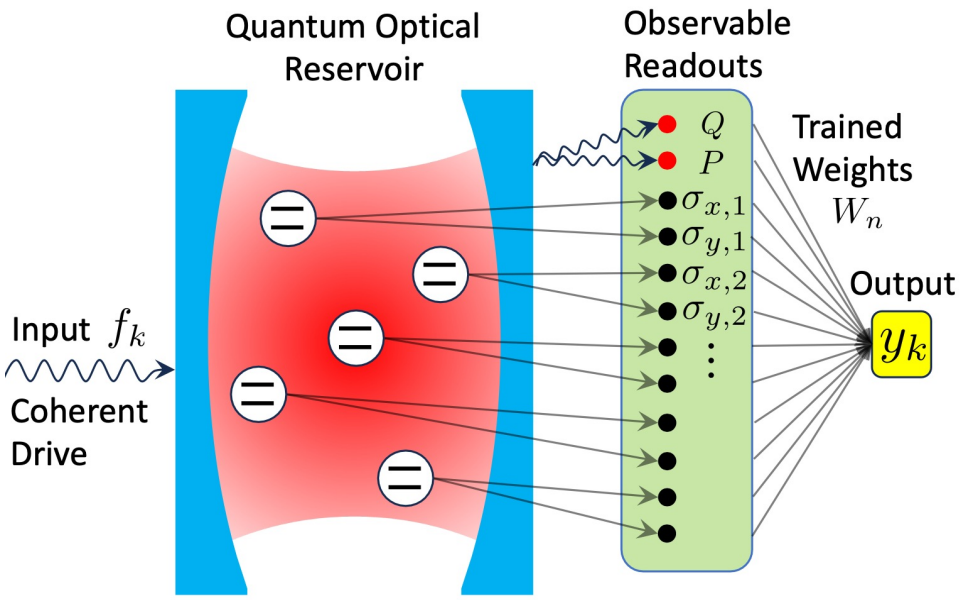
$$Q = c + c^\dagger, \quad P = i\left(c - c^\dagger\right);$$

$$\sigma_{x,i} = \sigma_i + \sigma_i^\dagger, \quad \sigma_{y,i} = i\left(\sigma_i - \sigma_i^\dagger\right)$$

[Zhu, C., Ehlers, P. J., Nurdin, H. I., & Soh, D. (2024). Practical and Scalable Quantum Reservoir Computing. arXiv preprint arXiv:2405.04799]

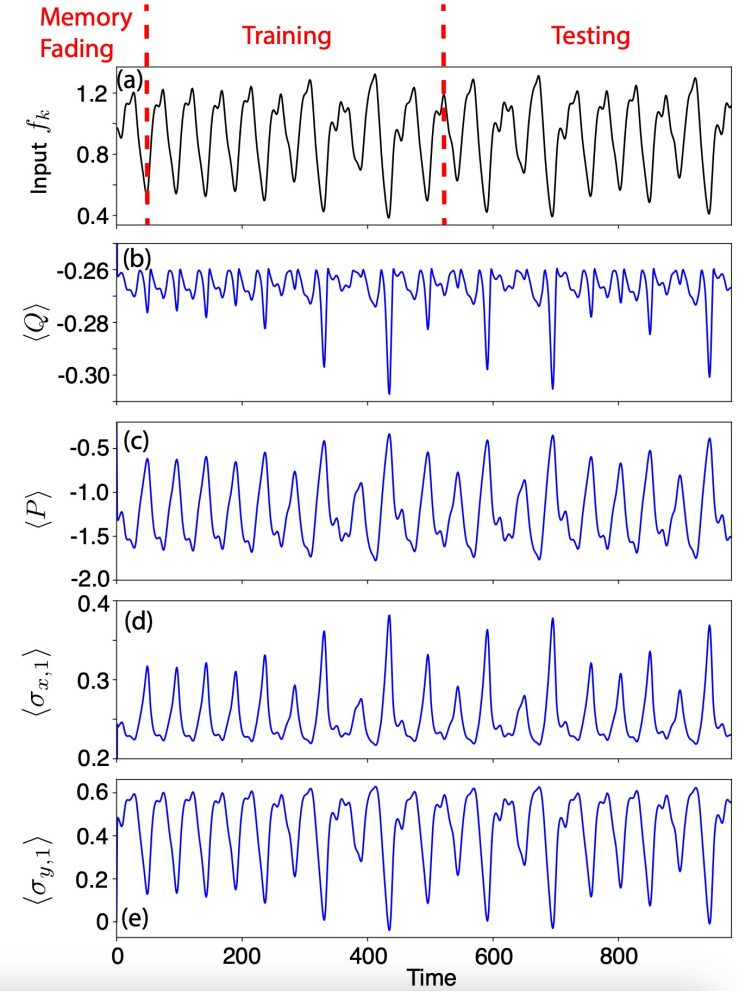
THE UNIVERSITY OF ARIZONA
Wyant College of Optical Sciences

Few-atom Reservoir Computing



(Continuous measurement with measurement back-action)

[Zhu, C., Ehlers, P. J., Nurdin, H. I., & Soh, D. (2024). Practical and Scalable Quantum Reservoir Computing. arXiv preprint arXiv:2405.04799]

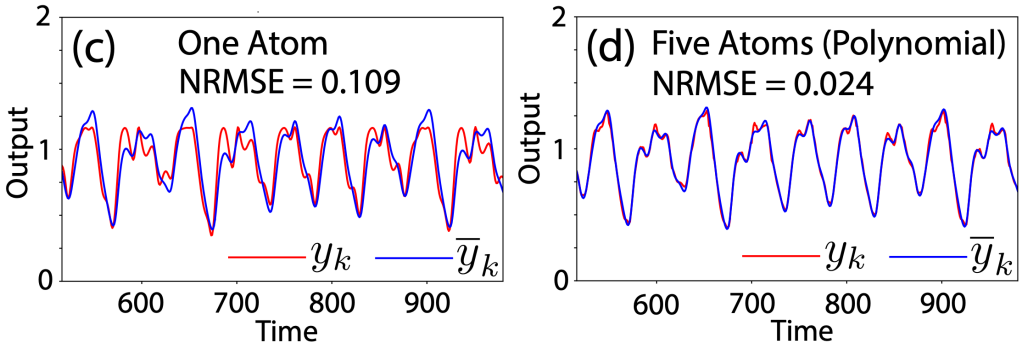


Mackey-Glass test

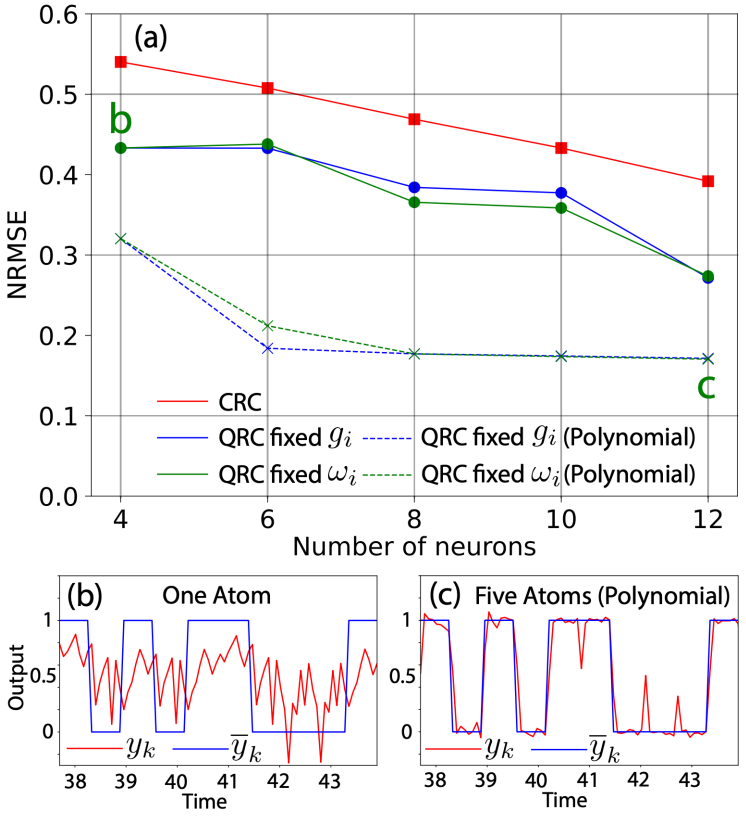
THE UNIVERSITY OF ARIZONA
Wyant College of Optical Sciences

Few-atom Reservoir Computing

- Mackey-Glass test – testing capability for predicting chaotic dynamic behavior



- Sine-Square classification test – testing nonlinear mapping capability

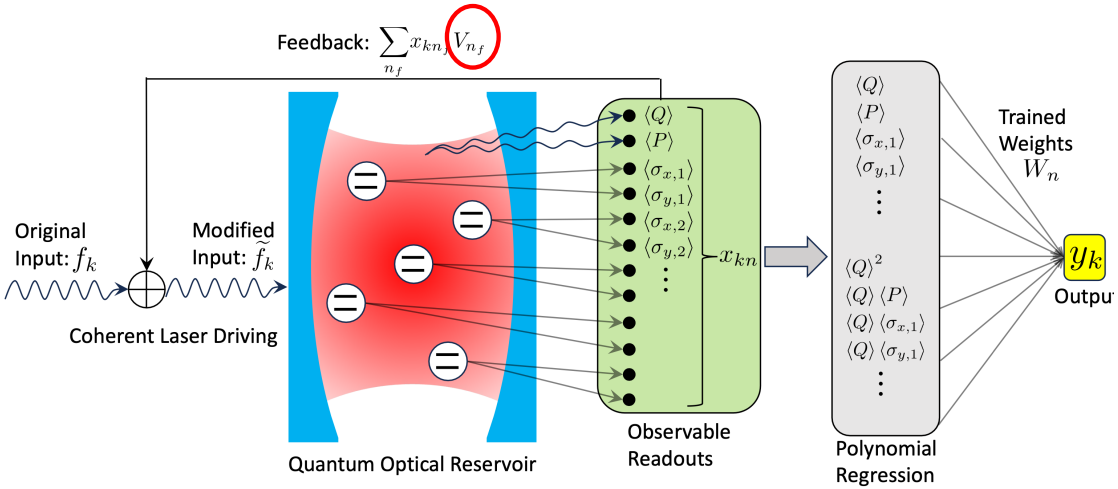


[Zhu, C., Ehlers, P. J., Nurdin, H. I., & Soh, D. (2024). Practical and Scalable Quantum Reservoir Computing. arXiv preprint arXiv:2405.04799]

THE UNIVERSITY OF ARIZONA
Wyant College of Optical Sciences

Can we do better with minimalistic quantum hardware?

- Making minimalistic quantum hardware much more complex by feedback



- We rigorously mathematically proved that feedback "**always**" improves the performance.

ESN without feedback:

$$x_{k+1} = g(Ax_k + Bu_k)$$

ESN with feedback:

$$x_{k+1} = g(Ax_k + B(u_k + V^T x_k))$$

Theorem 1 (*Superiority of Feedback for a Given ESN and Training Data*). *For any given matrix A and vector B in Eq. (8), and given sets of training inputs $\{u_k\} = \{u_k\}_{k=1,\ldots,N}$ and outputs $\{y_k\} = \{y_k\}_{k=1,\ldots,N}$ of finite length, define an optimized cost function $S_{\min}(A, B, \{u_k\}, \{y_k\})$ with appropriate optimal W and C. Then, for almost any given $(A, B, \{u_k\}, \{y_k\})$ except for vanishingly small number of $(A, B, \{u_k\}, \{y_k\})$, the feedback always reduces the cost function further:*

$$\min_V S_{\min}(A + BV^\top, B, \{u_k\}, \{y_k\}) < S_{\min}(A, B, \{u_k\}, \{y_k\}). \quad (23)$$

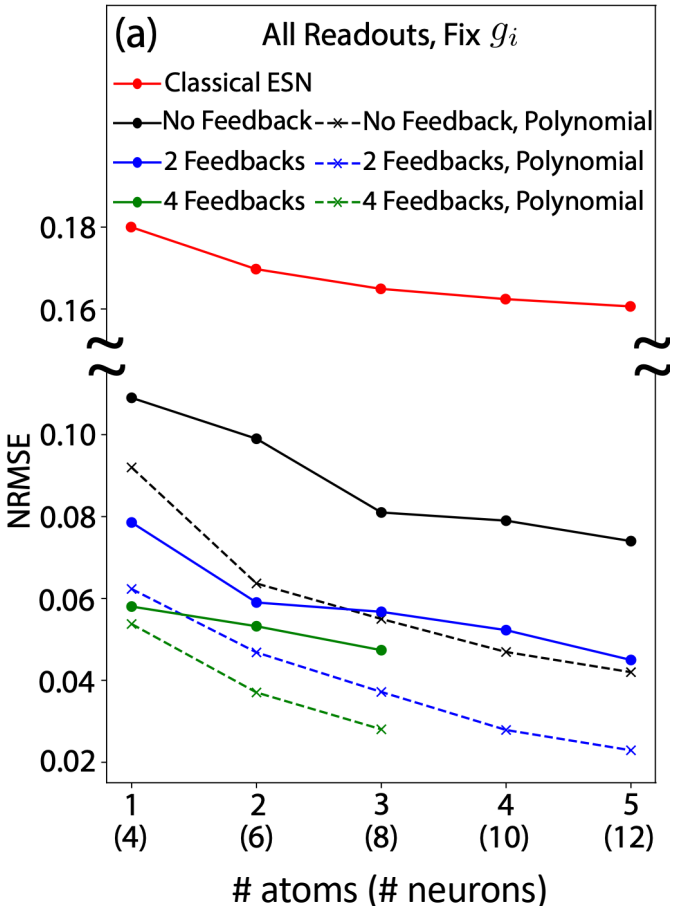
Moreover, if A is such that $A^\top A < a^2 \mathbb{I}_n$, where a is a constant that guarantees that the ESN is convergent, then the feedback gain V can always be chosen such that the ESN with feedback is also convergent and satisfy the above.

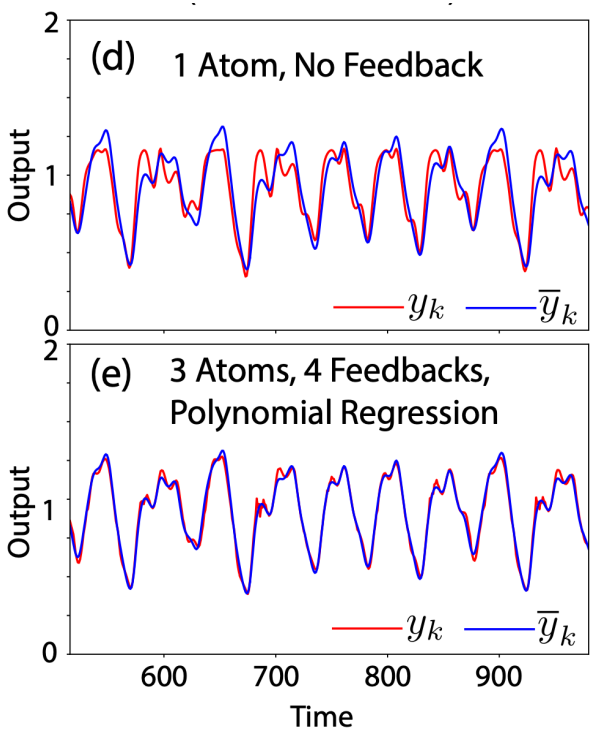
[Ehlers, P. J., Nurdin, H. I., & Soh, D. (2025). Improving the performance of echo state networks through state feedback. Neural Networks, 184, 107101.]

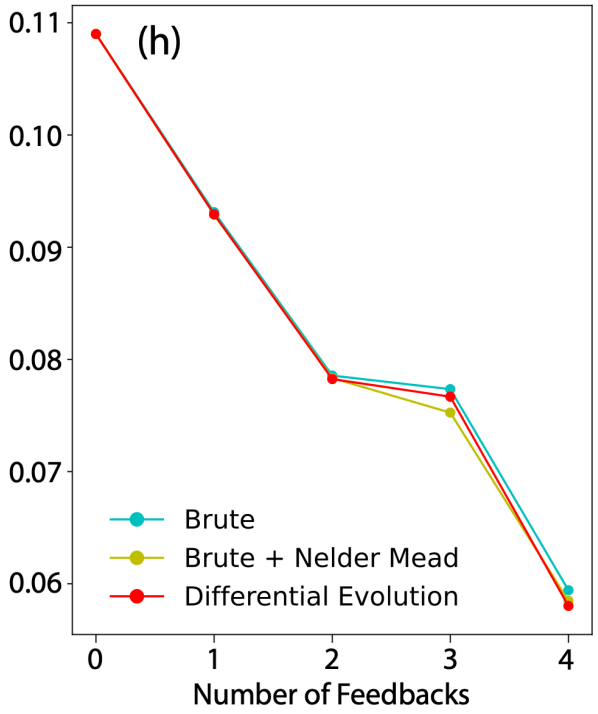
THE UNIVERSITY OF ARIZONA
Wyant College of Optical Sciences

Quantum Reservoir Computing with Feedback

- Mackey Glass test – testing the predicting capability for a chaotic dynamical behavior





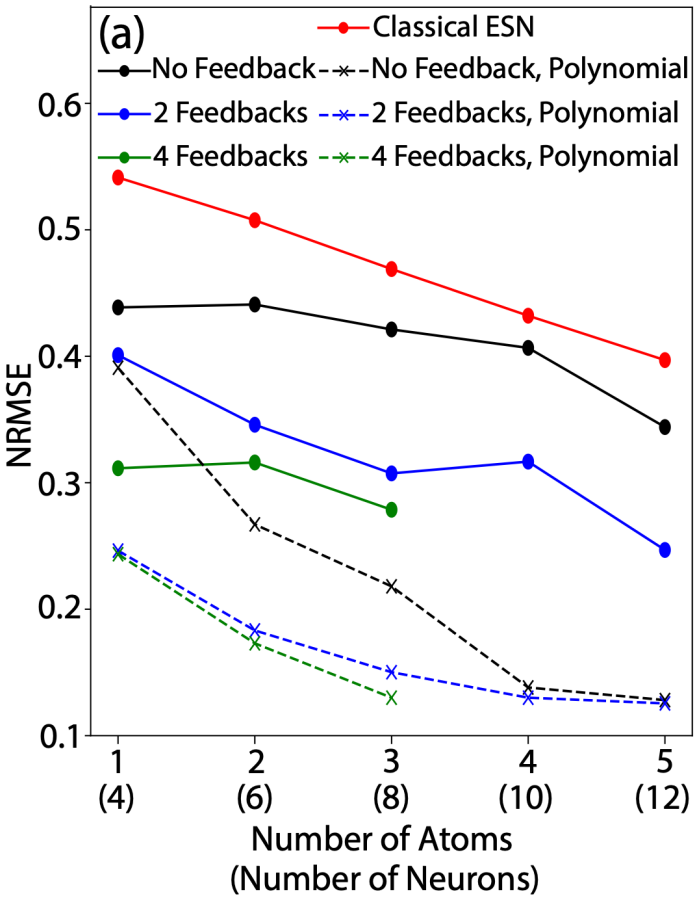


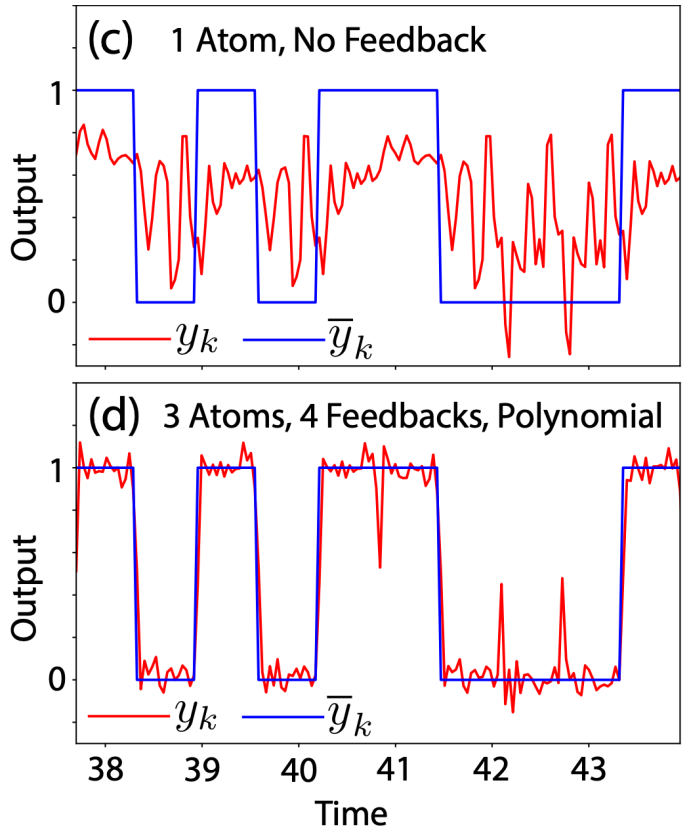
[Zhu, C., Ehlers, P. J., Nurdin, H. I., & Soh, D. (2024). Minimalistic and Scalable Quantum Reservoir Computing Enhanced with Feedback. arXiv preprint arXiv:2412.17817..]

THE UNIVERSITY OF ARIZONA
Wyant College of Optical Sciences

Quantum Reservoir Computing with Feedback

- Sine-Square classification test – testing nonlinear mapping capability



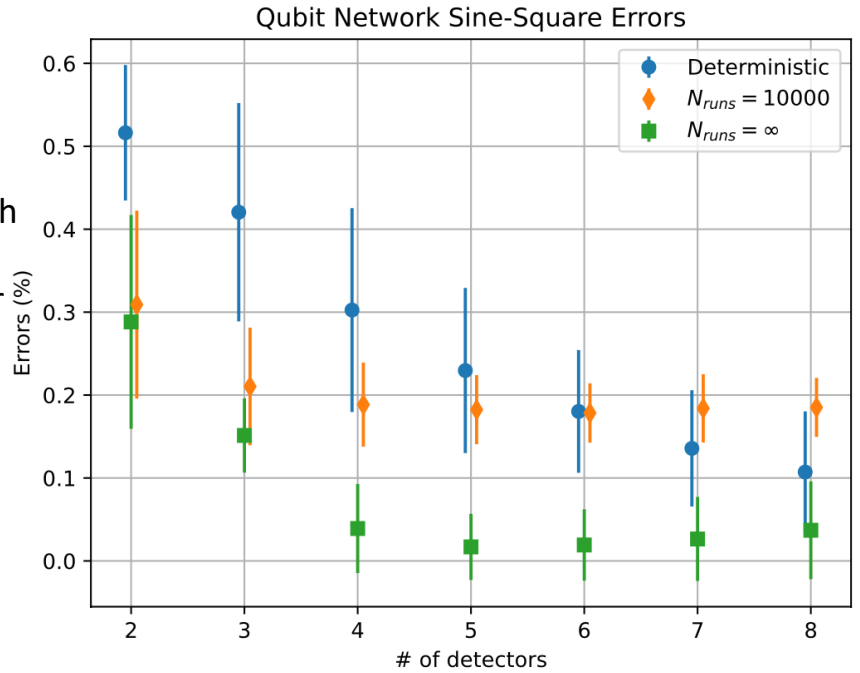


[Zhu, C., Ehlers, P. J., Nurdin, H. I., & Soh, D. (2024). Minimalistic and Scalable Quantum Reservoir Computing Enhanced with Feedback. arXiv preprint arXiv:2412.17817..]

THE UNIVERSITY OF ARIZONA
Wyant College of Optical Sciences

Forward looking – Stochastic Reservoir Computing: More Power Less Hardware

- **Key Idea**: Uses stochastic dynamical systems, leveraging probability distributions instead of fixed states.
- **Exponential Computational Scaling**
  - M stochastic nodes → 2^M computational states (exponential boost)
  - More computation with smaller hardware footprint
- **Compact Yet Powerful**
  - Qubit Reservoir Network: Uses quantum-inspired stochasticity for high efficiency
  - Stochastic Optical Network: Leverages photon detection for nonlinear transformations
  - **We proved universality through rigorous mathematics.**
- **Performance vs. Hardware Size**
  - Outperforms deterministic reservoirs when noise is low
  - Requires fewer physical nodes to achieve the same computation
- **Trade-offs**
  - Needs multiple runs to estimate probabilities
  - Shot noise can impact precision, but mitigable with more samples



[Ehlers, P. J., Nurdin, H. I., & Soh, D. (2024). Stochastic Reservoir Computers. arXiv preprint arXiv:2405.12382.]

THE UNIVERSITY OF ARIZONA
Wyant College of Optical Sciences

Summary

- New Paradigm 1 – Reservoir Computing
- New Paradigm 2 – Hardware Learning Machine
- New Paradigm 3 – Quantum Hardware Reservoir Computing

- Huge room for improvement via these new paradigm-shifting concepts

- The new paradigms apply broadly to
  - Both supervised/unsupervised learning
  - Knowledge distillation
  - Reinforcement learning
  - Chain of Thoughts
  - Distributed edge computing network

THE UNIVERSITY OF ARIZONA
Wyant College of Optical Sciences

Theory and Experiment of Scalable Quantum Systems Lab

Daniel Soh (PI)

Peter Ehlers (postdoc)

Chuanzhou Zhu (postdoc)

Andrew Pizzimenti (PhD student)

Charlotte Zehnder (PhD student)

Carter Gillenwater (PhD student)

Siddharth Vats (PhD student)

Ishwar Kaushik (MS student)

Phi Nguyen (PhD student)

Aamir Quraishy (PhD student)

Close collaborator – Dr. Hendra Nurdin, University of New South Wales, Australia

We welcome any form of collaborations with anyone!