***Titre de these:***

Le calcul par réservoir sur la plateforme de niobate de lithium sur isolant

***Mots clès***: calcul par réservoir, Ordinateur optique, L'informatique neuromorphique, photonique intégrée.

**Résumé**

Cette étude concerne le calcul par réservoir à retard temporel, en anglais Time-Delay Reservoir Computing (TDRC) dans les plateformes de photonique intégré, en particulier la plateforme Lithium Niobate On Insulator (LNOI). Nous proposons une nouvelle architecture intégrée « tout optique », avec seulement un déphaseur comme paramètre modifiable pouvant atteindre de bonnes performances sur plusieurs tâches de référence de calcul par réservoir. Nous étudions également l'espace de conception de cette architecture et le fonctionnement asynchrone du TDRC, qui s'écarte du cadre plus courant consistant à envisager les ordinateurs TDRC comme des réseaux. En outre, nous suggérons d'exploiter le schéma tout optique pour se passer du masque d'entrée, ce qui permet de contourner la conversion Optique/Electronique/Optique (O/E/O), souvent nécessaire pour appliquer le masque dans les architectures TDRC. Dans des travaux futurs, cela pourra permettre le traitement de signaux entrants en temps réel, éventuellement pour des applications de télécommunication de pointe. Les effets de la lecture électronique de sortie sur cette architecture sont également étudiés. Aussi, nous suggérons d'utiliser la corrélation de Pearson comme une métrique nous permettant de concevoir un réservoir capable de traiter plusieurs tâches en même temps sur le même signal entrant (et éventuellement sur des signaux dans des canaux différents).

Les premiers travaux expérimentaux menés à l'université RMIT sont également présentés. Par ces travaux, nous voulons étudier la performance de ces nouvelles architectures TDRC tout en ayant minimisant la complexité du matériel photonique. Pour cela on s’appuiera principalement sur les faibles pertes du LNOI qui permettent l'intégration du guide d'onde de rétroaction, et en utilisant uniquement l'interférence et la conversion d'intensité à la sortie (par le biais d'un photodétecteur) en tant que non-linéarité. Cela constitue une base sur laquelle pourront s’appuyer de futurs travaux étudiant les gains de performance lorsque des non-linéarités supplémentaires sont prises en compte (telles que celles de la plateforme LNOI) et lorsque la complexité globale du système augmente par l'introduction d'un plus grand nombre de paramètres.

Ces travaux portent donc sur l'exploration d'une approche informatique non conventionnelle particulière (TDRC), utilisant une technologie particulière (la photonique intégrée), sur une plateforme particulière (LNOI). Ces travaux s'appuient sur l'intérêt croissant pour l'informatique non conventionnelle puisqu'il a été démontré au fil des ans que les ordinateurs numériques ne peuvent plus être une solution unique, en particulier pour les applications émergentes telles que l'intelligence artificielle (IA). Le paysage futur de l'informatique englobera probablement une grande variété de paradigmes informatiques, d'architectures et de hardware, afin de répondre aux besoins d'applications spécialisées croissantes, tout en coexistant avec les ordinateurs numériques qui restent - du moins pour l'instant - mieux adaptés à l'informatique à usage général.

**PARTIE I : CONTEXTE SCIENTIFIQUE**

1. Introduction

L'essor de l'informatique au 20e siècle – partant des les matériaux de base jusqu'à l'échelle et l'étendue des applications - n'a pas été une simple percée technologique. Les progrès mathématiques du XIXe siècle avaient déjà transformé la logique philosophique en un langage formel qui a donné naissance à la logique symbolique, puis à l'algèbre de Boole, à la base de l'informatique numérique. En logique symbolique, une expression est évaluée pour sa valeur de vérité, produisant l'une des deux sorties possibles : Vrai ou Faux, qui peuvent en pratique être représentés comme les états « ON » et « OFF » d'un interrupteur. Les travaux pionniers de la théorie de la calculabilité dans les années 1930, catalysés par les travaux de Gödel sur la logique formelle, ont abouti à la machine de Turing : un ordinateur conceptuel qui manipule des symboles sur une longueur infinie de ruban suivant un tableau de règles. Cette machine simple peut mettre en œuvre, en principe, n'importe quel algorithme informatique. L'algèbre de Boole a également été largement utilisée par Claude Shannon pour développer la théorie de l'information, posant les bases de la théorie moderne de la communication.

D'un point de vue plus pratique, les progrès de la physique des semi-conducteurs et de la science des matériaux ont permis l'élaboration de la jonction p-n. Ainsi, l'informatique moderne n'a pu voir le jour que quand les sciences théoriques et pratiques se sont rencontrées au cours de la première moitié du XXe siècle.

Les ordinateurs à tube à vide ont été utilisés bien au-delà de la première moitié du XXe siècle. Cependant, ils étaient difficiles à entretenir en raison de pannes constantes (peu fiables) et fonctionnaient à des centaines de volts (gourmands en énergie). L'invention du transistor promettait (et plus tard réalisa) un changement d’état logique rapide, économe en énergie et fiable. Cela a permis aux ordinateurs d'augmenter leur nombre d’interrupteurs (grâce à la réduction de la taille des dispositifs et à l'intégration sur puce) et d'élargir le champ des applications, leur permettant d’être les machines performantes et omniprésentes que nous connaissons aujourd'hui.

L'incroyable succès de la microélectronique , stimulé par la loi de Moore, prévoyant que le nombre de transistors sur un circuit intégré doublerait environ tous les deux ans, a soutenu l'amélioration des performances informatiques pendant plus de 50 ans. Notons également les progrès continus au niveau des architectures des dispositifs permettant de réduire la taille des transistors à 2 nm (y compris Fin-Fet, GAAFET, etc.), mais au prix d'une augmentation des coûts de R&D et de fabrication. En outre, les ordinateurs d'aujourd'hui - qui utilisent essentiellement l'architecture connue sous le nom d'architecture « Princeton » ou plus communément de modèle « von Neumann » - souffrent de ce qu’on appelle un goulot d’étranglement.

Celui-ci se situe au niveau du transfert de données entre les unités distinctes de mémoire et de calcul. C'est pourquoi l'exploration de paradigmes informatiques non conventionnels a suscité un intérêt croissant, allant du calcul en mémoire (les opérations de calcul et la mémoire se situent dans la même unité), au calcul stochastique, en passant par le calcul neuromorphique et le calcul par réservoir, pour ne citer qu’eux. Le calcul optique a également été utilisé comme un moyen d'exploiter les avantages à travailler avec la lumière plutôt qu’avec des signaux à fréquence radio (RF), ou simplement comme base pour la mise en œuvre de certains des paradigmes susmentionnés.

La photonique peut faire la différence dans le domaine de l'informatique, notamment en termes de :

- **Largeur de la bande de fréquence** : la photonique fonctionne avec des ondes lumineuses (des centaines de THz) alors que l'électronique numérique fonctionne avec des ondes RF (GHz). La différence de largeur de bande atteint alors 5 ordres de grandeur. Un parallélisme massif peut être obtenu grâce au multiplexage des fréquences, les peignes en fréquence pouvant générer de la lumière dans environ lignes de fréquence distinctes ; les données de chaque ligne pouvant alors être traitées en parallèle.

- **Dynamique à faible perte:** les photons peuvent se propager avec des pertes négligeables dans les configurations en espace libre et des pertes très faibles dans certaines plates-formes intégrées (par exemple, les pertes démontrées sur les plates-formes LNOI ou avec SiN sont de l'ordre de 0,2 dB/cm dans la bande optique C). C’est une autre raison pour laquelle, on s'attend à ce que les interconnexions optiques jouent également un rôle important dans la communication électronique intra-puce.

L'utilisation de plates-formes photoniques pour des applications en calcul est donc très prometteuse, en particulier pour les paradigmes émergents qui promettent une augmentation de la vitesse et de l'efficacité du calcul, par exemple dans le domaine du traitement du signal ou de l'intelligence artificielle.

1. Quelques éléments de photonique

Cette section présente un bref aperçu de la photonique, en mettant l'accent sur les circuits intégrés photoniques (PICs). De nombreuses recherches ont été menées pour trouver des matériaux adaptés aux dispositifs et circuits photoniques intégrés. Pour la plupart des plates-formes, la photolithographie est utilisée de la même manière qu’en microélectronique, pour la gravure et le dépôt de matériaux.

Mais, contrairement aux circuits électroniques où les électrons existent facilement à l'intérieur des matériaux, les circuits photoniques ont besoin de sources de lumière, généralement obtenus par émission stimulée en générant des paires électron-trou dans un matériau à gain. La lumière a également besoin d’être détectée, la détection se faisant dans un semiconducteur par la combinaison de ces paires électron-trou.

D'un point de vue historique, la conception des dispositifs a précédé de quelques décennies les réflexions au niveau du circuit ou de systèmes. Cela a entraîné la spécialisation des plateformes pour ces différents dispositifs, par exemple : le phosphure d'indium (InP) pour les lasers, l'arséniure de gallium (GaAs) pour les détecteurs et le niobate de lithium LiNbO3 pour les modulateurs à grande vitesse. Chacun de ces dispositifs a été fabriqué selon des procédés de fabrication différents qui sont fondamentalement incompatibles avec ceux utilisés pour les autres dispositifs.

Si on se focalise maintenant au niveau de la conception et de la simulation des dispositifs et des systèmes photoniques il est nécessaire de revenir à l’optique fondamental. La physique de la lumière peut être décrite par différentes théories, qui suivent l'évolution de notre compréhension de la nature de la lumière. Par ordre chronologique, ces théories sont l'optique géométrique, l'optique des ondes scalaires, l'optique des faisceaux, l'optique de Fourier, l'optique électromagnétique et l'électrodynamique quantique. Ces différentes théories de la lumière peuvent être utilisées pour modéliser la propagation de la lumière dans les circuits photoniques, selon que l'on souhaite une simulation au niveau du dispositif ou au niveau du système, et selon les types de phénomènes pris en compte. Pour les implémentations photoniques intégrées au niveau du système, on modélise la plupart du temps la lumière grâce à l’optique ondulatoire à ondes guidées (par exemple, la théorie des modes couplés). Les simulations au niveau du circuit effectuées dans cette thèse sont réalisées à l'aide de la méthode de la matrice S. Un exemple simple est présenté ci-dessous pour montrer comment cela fonctionne.

Considérons une source laser, décrite comme un champ électromagnétique cohérent et monochromatique. Le champ électrique de cette source de fréquence, en supposant une onde plane (c'est-à-dire sans dépendance spatiale), peut être décrit par son amplitude *A* et sa phase initiale comme suit :

Cette lumière peut être couplée à une puce photonique intégrée par divers moyens, le plus souvent par l'intermédiaire d'un coupleur à réseau. Lors du couplage, le champ est donné par :

où est le coefficient de couplage du CG qui dépend de la longueur d'onde du laser et de l'angle entre la fibre et la normale. Après avoir été couplé dans le guide d'onde de la puce, il se propage sur une certaine distance, ce qui donne:

où est la longueur du guide d'ondes [m], est la perte du guide d'ondes [], et [] est la constante de propagation du mode guidé avec l'indice de réfraction effectif . L'indice effectif décrit en première approximation l'indice que le champ subit lors de sa propagation dans le guide d’ondes sachant qu’il ne se déplace pas uniquement dans le cœur du guide. Cela signifie qu'une partie du mode s'échappe dans le milieu environnant, se faisant, le champ rencontre un indice « effectif » qui tient compte de cet effet. D'autres éléments composants les circuit photoniques peuvent être décrits de manière similaire qu'il s'agisse d'éléments actifs, passifs, linéaires ou non-linéaires. Nos simulations sont réalisées dans le domaine temporel en raison de l’aspect dynamique du calcul TDRC.

A propos du calcul par réservoir

Comme le calcul neuromorphique , le calcul par réservoirs (RC) est un type d'approche d'apprentissage automatique supervisé. Initialement connu sous le nom de « Echo-state networks » (ESN), le RC a été utilisé dans des implémentations logicielles au début des années 2000 pour contourner les problèmes de formation de grands réseaux neuronaux récurrents (RNN). Différents systèmes physiques peuvent être exploités pour mettre en œuvre les RC, notamment dans les domaines de l'électronique et de la photonique. L'ESN est un système dynamique non linéaire qui présente des caractéristiques particulières, à savoir

- la propriété d'écho-état (ESP) ; le système permet des réverbérations d'entrées passées qui diminuent avec le temps (mémoire déclinante),

- l'expansion de la dimensionnalité, généralement obtenue grâce à une fonction d'activation non linéaire.

L'ESP est une condition qui nécessite l'oubli des entrées passées, de sorte que le réservoir ne dépende pas de ses conditions initiales. Il s'agit d'une caractéristique essentielle du RC car elle permet au réservoir de répondre de la même manière à la même entrée. Cette mise en correspondance cohérente est essentielle pour qu'un modèle formé sur la sortie du réservoir puisse se généraliser. L'expansion de la dimensionnalité permet de séparer plusieurs classes avec de simples classificateurs linéaires, même pour des tâches qui nécessiteraient normalement des réseaux neuronaux plus vastes et plus profonds. L'idée principale est qu'en projetant les données dans un espace de dimension supérieure, il existe davantage d'hyperplans capables de bien séparer les différentes classes de données, ce qui permet d'obtenir une précision élevée.

Tout schéma de RC se compose des trois couches suivantes :

- La couche d'entrée : où les données sont potentiellement prétraitées puis injectées dans le réservoir,

- La couche réservoir : composée de nœuds non linéaires connectés les uns aux autres,

- La couche de sortie : où les états du réservoir sont collectés au moyen d'un mécanisme de lecture.

Chaque couche est reliée à la suivante par une série de connexions pondérées. La distinction entre l'informatique à réservoir des réseaux neuronaux temporels similaires, tels que les RNN, réside dans le fait que les poids entre la couche d'entrée et la couche de réservoir peuvent être définis et fixés, de même que les interconnexions entre les nœuds non linéaires, alors que dans les RNN, toutes les connexions doivent être optimisées. Ainsi, dans le RC, le système s’entraîne uniquement sur la couche de sortie, ce qui signifie que seules les méthodes des moindres carrés sont nécessaires pour trouver les poids optimaux.

Le RC peut être réalisé spatialement ou temporellement, dans le cas des TDRC, si le réseau consiste en des multiplexés dans le temps. Dans ce dernier cas, il s'agit d'utiliser un seul nœud non linéaire et de l'échantillonner dans le temps pour obtenir les états souhaités du nœud.

Notre travail s’est alors focalisé sur la conception de ces TDRC utilisant des circuits photoniques. Ce sujet a été exploré depuis 2011, et un chapitre détaillé sur l'état de l'art aborde les innovations les plus notables dans ce domaine.

**Partie II : Contributions scientifiques**

1. Calcul par réservoir photonique de complexité minimale

Notre première contribution s'appuie sur la littérature existante et sur le thème contemporain du calcul par réservoir dans les circuits intégrés photoniques, plutôt qu'en espace libre. Notre étude montre comment un dispositif simple (illustré sur la Fig. 1), composé de guides d'ondes passifs et d'un déphaseur chauffant (pour la reconfigurabilité), peut déjà être aussi performant que certains schémas plus complexes. La non-linéarité se retrouve alors dans la non-linéarité de conversion d'intensité fournie au photodétecteur. Les performances ont été mesurées sur une variété de tâches standard de référence.

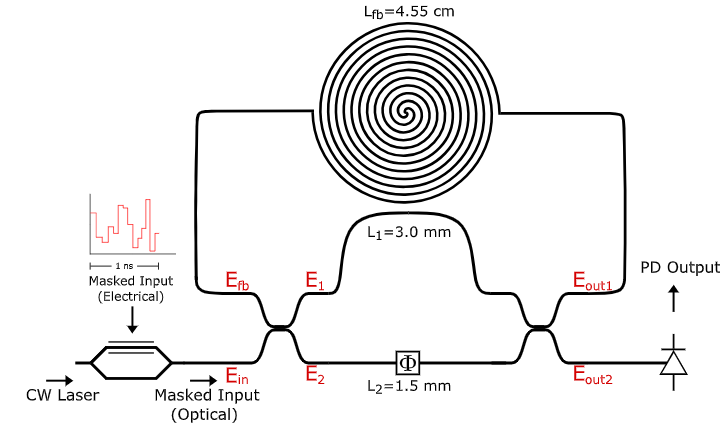


Fig. 1: Architecture de calcul par réservoir photonique reconfigurable de complexité minimale.

Pour ce qui est de son fonctionnement, l'élément chauffant modifie la longueur du chemin optique dans le guide d'ondes inférieur, ce qui entraîne une modification de la dynamique de sortie. Une fois la dynamique souhaitée trouvée, le circuit peut être entraîné à résoudre des tâches de référence avec une grande précision, telles que les tâches XOR bit à bit, Mackey-Glass et Santa-Fe.

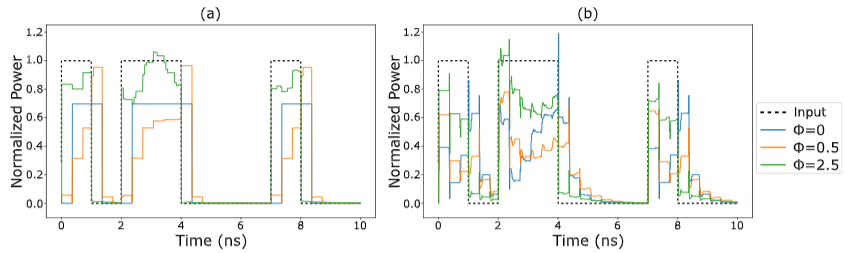


Fig. 2: Sortie du système pour différents déphasages (tension de chauffage)

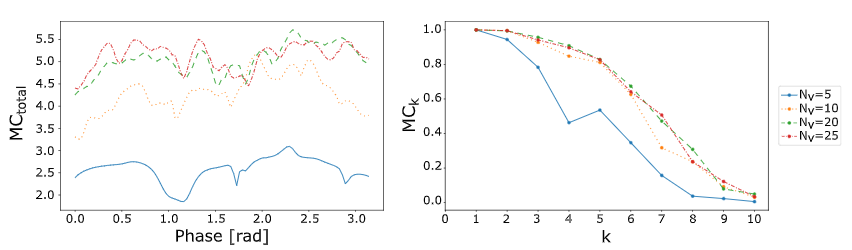


Fig. 3: (a) Capacité totale de la mémoire et (b) capacités individuelles maximales

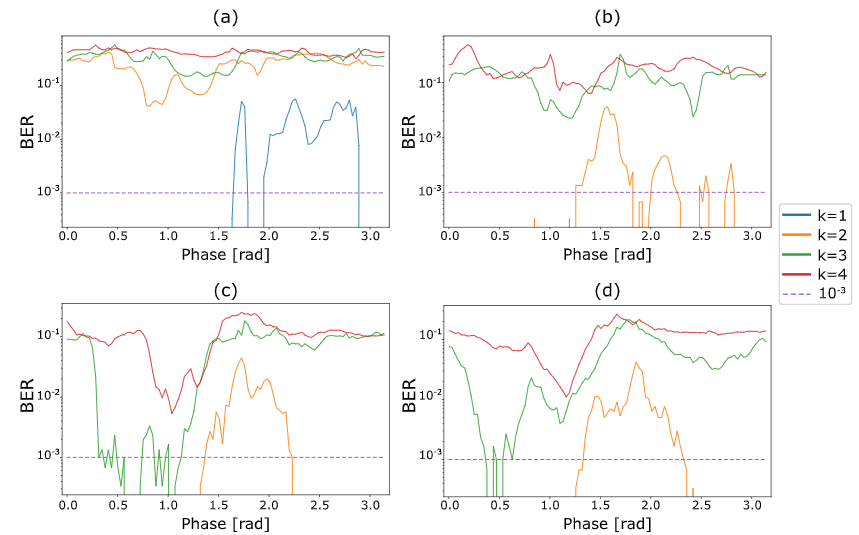


Fig. 4: Performances du XOR pour différentes profondeurs de mémoire, sous différentes valeurs de déphasage.

Nous avons également montré qu’il était possible que la rétroaction soit intégrée à la puce, grâce aux faibles pertes de la plate-forme LNOI. Cette étude montre en fin de compte ce qu'un simple schéma TDRC peut faire, et servira de référence lorsque des schémas plus complexes seront envisagés.

1. Calcul par réservoir photonique asynchrone sans masque

L'une des principales limites du TDRC se situe dans le multiplexage temporel des nœuds limitant la vitesse de traitement du système. Cette limitation est généralement accentuée par l'utilisation d'un masque d'entrée, qui réduit encore la vitesse de traitement effective du système. Ce masque empêche le système de tomber dans un état stable pendant la durée de l'échantillon d'entrée. De plus, ce masque ajoute des contraintes en termes de complexité et de consommation d'énergie puisqu'une conversion O/E/O peut être nécessaire pour appliquer le masque. Dans notre étude, nous explorons la possibilité d'utiliser le réservoir sans le masque d'entrée en utilisant la longueur du guide de rétroaction de telle manière à remplir le même rôle que le masque. Le système n’a alors plus besoin de masque d'entrée accélérant ainsi son fonctionnement. Nous avons testé le réservoir sur une série de mesures ne dépendant pas de de la tâche choisie, ainsi que sur quelques tâches de référence. Les résultats numériques montrent de bonnes performances sur ces différentes tâches de référence, comparable à d'autres schémas dans la littérature (utilisant un masque).

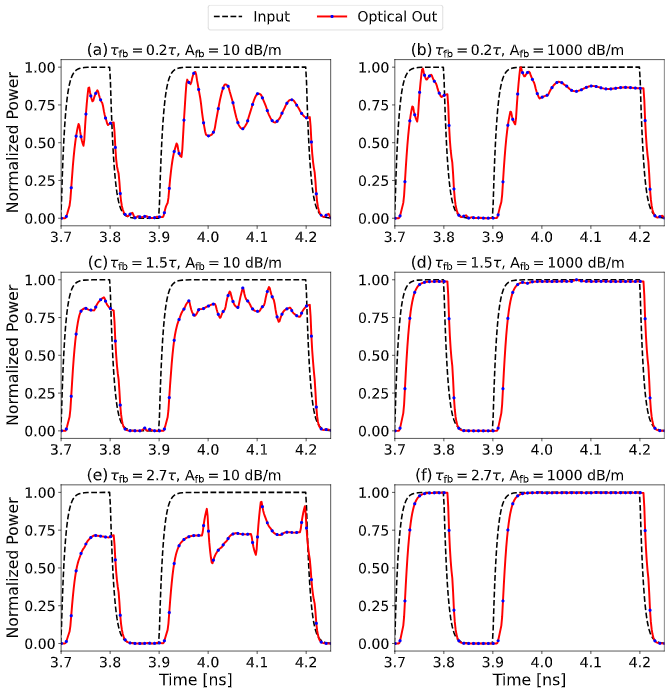


Fig. 5: Des longueurs de rétroaction plus courtes permettent au signal d'interagir avec lui-même plusieurs fois, ce qui permet d’avoir le même effet qu’un masquage de l'entrée. Cela étant possible grâce aux faibles pertes de la plateforme LNOI.

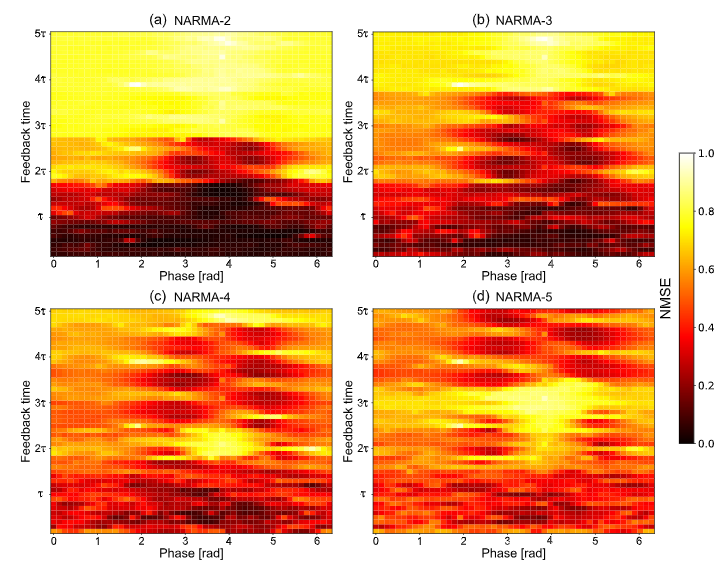


Fig. 6: L'erreur quadratique moyenne normalisée (NMSE) est présentée pour une variété de tâches avec différents temps de rétroaction et déphasages.

1. Effet d'une lecture à largeur de bande limitée

Nous examinons également les effets de la lecture électronique sur les performances de cette architecture. La lecture joue un rôle central, étant l'intermédiaire entre le réservoir optique et le signal électronique de sortie. La lecture, composée d'une photodiode et des circuits associés, imprime des imperfections sur le signal qui sera utilisé pour l'apprentissage du modèle des moindres carrés. Pour cela, nous considérons une photodiode simple et un modèle de bruit, et nous tenons compte de la conversion analogique-digital (ADC) effectuée par l'oscilloscope numérique. Pour l'ADC nous avons utilisé une seule acquisition avec un nombre effectif de bits inférieur pour vérifier que le schéma peut fonctionner dans ces conditions pour certaines tâches. Les ENOB considérés sont généralement utilisés dans les oscilloscopes haut de gamme.



Tableau 1 : Le système est capable de résoudre certaines tâches même en considérant une routine d'acquisition unique avec un nombre effectif de bits inférieur.

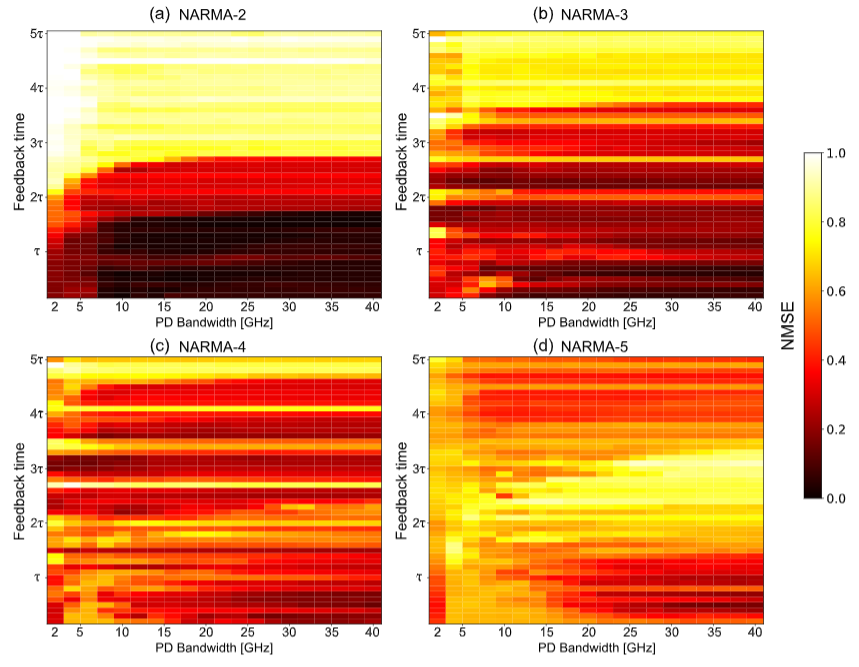


Fig. 7: Les performances sont enregistrées pour différentes largeurs de bande du photodétecteur et différents temps de rétroaction, montrant une grande tolérance pour des largeurs de bande plus faibles pour certaines tâches.

1. Démonstration expérimentale

Dans cette dernière étude, nous avons fabriqué et testé expérimentalement notre schéma de calcul par réservoir à complexité minimale sans masque sur la tâche de référence Santa-Fe, en obtenant d'excellents résultats. La puce a été fabriquée par l'équipe de fabrication du Centre intégré de photonique et d'applications du RMIT. Le design de la puce a été réalisé à l'aide d'IPKISS et du kit de développement de processus (pdk) d'InPAC. Différentes règles de conception ont été vérifiées manuellement pour s'assurer que :

- Un minimum de courbes se chevauchant avec les lignes de la grille de lithographie par faisceau d'électrons (EBL) ;

- la distance minimale entre les guides d'ondes soit respectée ;

- L'espacement des plots DC (200 m) corresponde au pas de la sonde DC ;

- L'espacement des contacts RF corresponde au pas de la sonde RF.

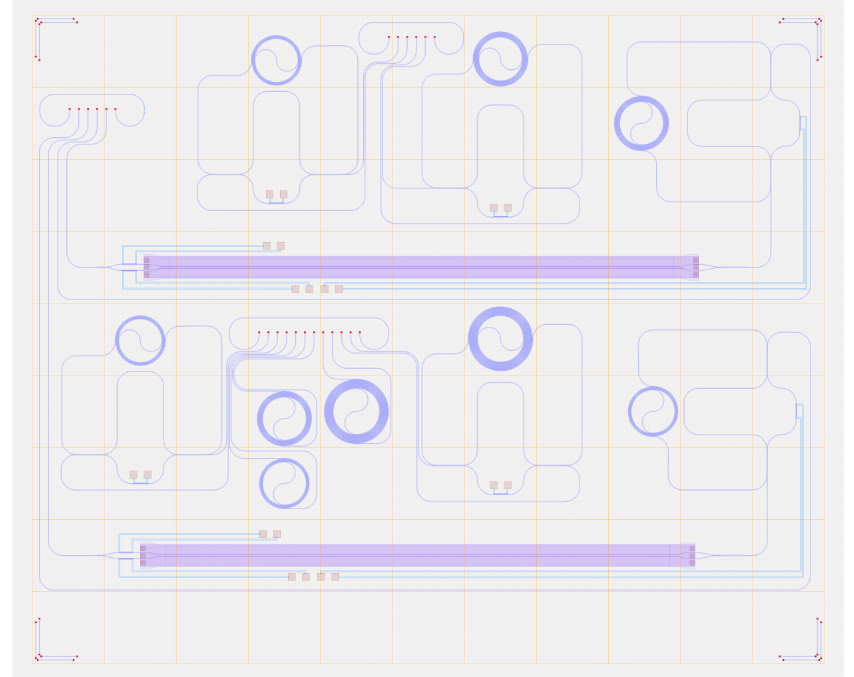


Fig. 8: Disposition de la puce générée à l'aide d'IPKISS. Plusieurs dispositifs de test ont été réalisés avec des longueurs de rétroaction variables, ces dispositifs possèdent des modulateurs intégrés, ainsi que des structures de test.

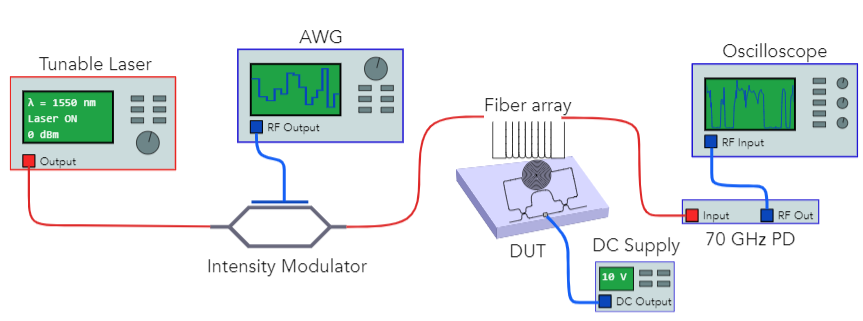


Fig. 9: Dispositif expérimental

Les résultats expérimentaux de la tâche de Santa Fe sont présentés sur la Fig.10 pour tous les ensembles de données obtenues par balayage en longueur d'onde et en tension au déphaseur. Nous avons considéré deux dispositifs, celui utilisant une boucle de rétroaction de 2,84 cm (R284) et celui utilisant une boucle de rétroaction de 1,88 cm (R188). Normalement, la tâche de Santa Fe est considérée pour la prédiction à un pas, mais nous avons également documenté les résultats pour la prédiction à 2 pas et à 3 pas de temps. On peut constater que R188 est plus performant que R284, en raison des pertes de rétroaction plus faibles rencontrées.

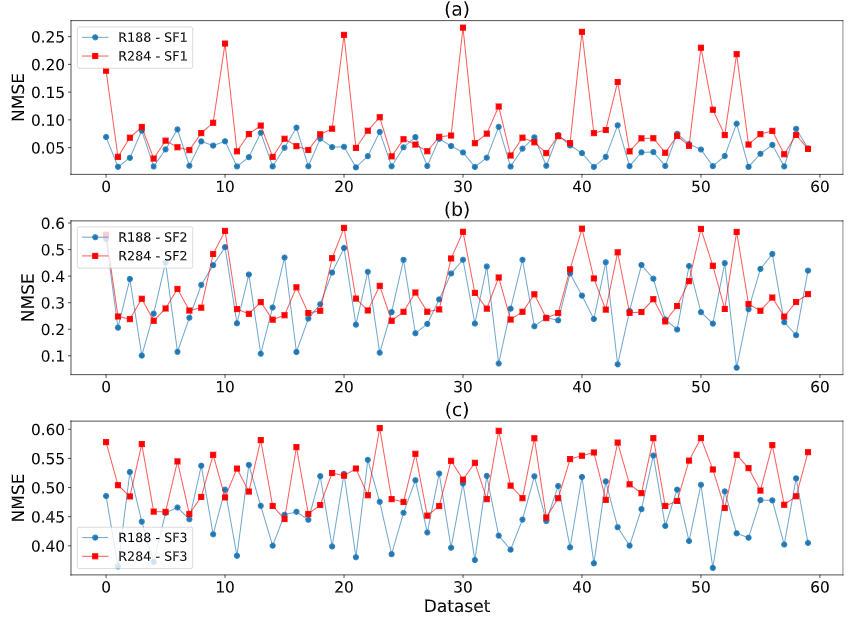


Fig. 10 : NMSE obtenu avec les différents ensembles de données pour la tâche de Santa Fe pour (a) la prédiction 1 pas, (b) la prédiction 2 pas, et (c) la prédiction 3 pas.

1. Résumé et conclusions

Au cours de ces travaux nous avons cherché à minimiser la complexité des schémas de RC photonique et avons obtenus de bons résultats. Pour cela, nous avons lancé plusieurs recherches. La première consiste à proposer une nouvelle architecture de RC photonique intégrée qui est reconfigurable à l'aide d'un seul élément de déphasage, qui contrôle deux paramètres simultanément : la force de rétroaction et la phase de rétroaction. Se faisant, la modification de la puissance de la rétroaction ne nécessite pas de couplage de puissance hors du système, comme c'est généralement le cas en utilisant un atténuateur optique. Les plateformes à faible perte (LNOI, SiN) permettent l'intégration d’un guide d'onde de rétroaction. Notre conception simplifiée a permis d'obtenir de bonnes performances pour certaines tâches de référence en matière de RC, tout en réduisant considérablement la complexité matérielle du réservoir et de l'équipement auxiliaire.

La deuxième étude porte sur une réduction supplémentaire de la complexité en éliminant le masque d'entrée au niveau de la couche d'entrée et en concevant la longueur de la rétroaction pour obtenir une dynamique qui reste transitoire pendant la perturbation de l'entrée, ce qui était essentiellement l'objectif du masque d'entrée. Les performances du réservoir dans le régime asynchrone ont été évaluées à l'aide d'une variété de mesures (indépendantes de la tâche effectuée mais également sur de tâches de référence). Les résultats ont montré des performances comparables à celles de la littérature pour les tâches étudiées.

En résumé, les résultats présentés dans cette thèse suggèrent que de tels schémas RC minimaux peuvent atteindre des performances comparables à celles des schémas qui nécessitent des instruments auxiliaires coûteux ou des schémas qui utilisent une multitude de paramètres. Enfin, les effets de la lecture à bande passante limitée ont été étudiés, en montrant comment le bruit et la limitation de la bande passante affectent les performances du système pour diverses tâches. Nous avons conclu que la nature de la tâche détermine la robustesse des performances du système pour des largeurs de bandes plus faibles. Enfin, la démonstration expérimentale a été présentée et le système a permis de résoudre la tâche de référence de Santa Fe avec d'excellents résultats.

L'architecture proposée présente plusieurs limites essentielles, en particulier liées aux échelles de temps. La première est la nécessité d'un équipement de lecture à grande vitesse pour atteindre des largeurs de bande plus élevées. Cela implique des photodétecteurs et des oscilloscopes à large bande passante, qui sont coûteux et encombrants.

Comme d'autres systèmes basés sur des puces, il est limité à la résolution de tâches requérant une mémoire moyenne ou faible. Par exemple, il est très difficile de résoudre la tâche NARMA-10 avec ce système, surtout si l'on tient compte du bruit de lecture. En outre, bien que nous ayons abordé la question de la résolution de plusieurs tâches en même temps, des recherches plus approfondies sont encore nécessaires. Il serait notamment intéressant d'étudier l'utilisation de plusieurs longueurs d'onde, comme on pourrait le faire en utilisant une source de peigne en fréquence intégrée, ce qui est possible sur la plateforme LNOI. Aussi, il serait intéressant d’exploiter la grande variété de non-linéarités optiques accessibles sur le LNOI pour créer des systèmes de RC et comparer leur performance au schéma minimaliste.

Enfin, et cela n'est pas spécifique à notre design, les systèmes photoniques souffrent intrinsèquement de décalages de la longueur d’onde due aux changements d'indice de réfraction provoqués par les changements de température. S'il est possible de compenser ce phénomène en laboratoire, cela n'est pas évident dans un environnement conditionné pour des applications réelles. Bien que plusieurs études aient exploré les moyens d'atténuer ce problème, des améliorations sont encore possibles. L'un des moyens de résoudre ce problème à l'aide de l'architecture proposée pourrait être l'utilisation d'un FPGA ou d'un ASIC cointégré qui met en œuvre un algorithme d'étalonnage pour compenser la dérive de la température. L'ASIC prendrait en entrée une partie de la puissance de sortie détectée et ajusterait en conséquence la tension du déphaseur pour effectuer la compensation.

***Thesis Title***:

Reservoir Computing in Lithium Niobate on Insulator Platforms

***Keywords***: reservoir computing, photonic computing, neuromorphic computing, integrated photonics.

**Abstract**

This work concerns time-delay reservoir computing (TDRC) in integrated photonic platforms, specifically the Lithium Niobate on Insulator (LNOI) platform. We propose a novel all-optical integrated architecture, which has only one tunable parameter in the form of a phase-shifter, and which can achieve good performance on several reservoir computing benchmark tasks. We also investigate the design space of this architecture and the asynchronous operation, which represents a departure from the more common framework of envisioning time-delay reservoir computers as networks in the stricter sense. Additionally, we suggest to leverage the all-optical scheme to dispense with the input mask, which allows the bypassing of an O/E/O conversion, often necessary to apply the mask in TDRC architectures. In future work, this can allow the processing of real-time incoming signals, possibly for telecom/edge applications. The effects of the output electronic readout on this architecture are also investigated. Furthermore, it is suggested to use the Pearson correlation as a simple way to design a reservoir which can handle multiple tasks at the same time, on the same incoming signal (and possibly on signals in different channels). Initial experimental work carried out at RMIT University is also reported. The unifying theme of this work is to investigate the performance possibilities with minimum photonic hardware requirements, relying mainly on LNOI’s low losses which enables the integration of the feedback waveguide, and using only interference and subsequent intensity conversion (through a photodetector) as the nonlinearity. This provides a base for future work to compare against in terms of performance gains when additional nonlinearities are considered (such as those of the LNOI platform), and when overall system complexity is increased by means of introducing more tunable parameters. Thus, the scope of this work is about the exploration of one particular unconventional computing approach (reservoir computing), using one particular technology (photonics), on one particular platform (lithium niobate on insulator). This work builds on the increasing interest of exploring unconventional computing, since it has been shown over the years that digital computers can no longer be a `one-size-fits-all', especially for emerging applications like artificial intelligence (AI). The future landscape of computing will likely encompass a rich variety of computing paradigms, architectures, and hardware, to meet the needs of rising specialized applications, and all in coexistence with digital computers which remain --- at least for now --- better suited for general-purpose computing.

**PART I: SCIENTIFIC CONTEXT AND BACKGROUND**

1. Introduction

The rise of computing in the 20th century --- from enabling materials to the scale and breadth of applications --- was not simply a technological breakthrough. Mathematical progress in the 19th century had already transformed philosophical logic into a formal language which gave birth to symbolic logic, and subsequently Boolean algebra, which is the foundation of digital computing. In symbolic logic, an expression is evaluated for its truth value, yielding one of two possible outputs: True or False, which can in practice be represented as 'ON' and 'OFF' states of a switch. Pioneering work in computability theory in the 1930s, catalyzed by Gödel's then-contemporary work in formal logic, led to the Turing machine: a conceptual computer that manipulates symbols on an infinite length of tape according to a table of rules. This simple machine could implement, in principle, any computer algorithm. Boolean algebra was also heavily utilized by Claude Shannon to develop Information Theory, laying the groundwork for modern communication theory. On the more practical side, advancements in semiconductor physics and material science enabled the engineering of the p-n junction. Thus, modern computing could only emerge thanks to the successive breakthroughs spanning the theoretical and practical sciences in the first half of the 20th century.

Vacuum tube computers were being used well beyond the first half of the 20th century. However they were difficult to maintain due to constant failures (unreliable), and were also operated at hundreds of volts (power hungry). The invention of the transistor promised (and later delivered) rapid, energy-efficient, and reliable switching. This allowed for the scaling up of computing in terms of number of switches (through scaling down of device size and on-chip integration) and also the scope of applications, making them the ubiquitous machines we have today. The incredible success of digital electronics, driven by Moore's law, which predicted that the number of transistors on an integrated circuit roughly doubles every two years, sustained the improvement of computing performance for more than 50 years. Continuous progress in device architectures has allowed further scaling down to 2-nm process (including Fin-Fet, GAAFET, etc.. however at the expense of rising R&D and fab costs. Furthermore, today's computers --- which essentially use the architecture known as the ``Princeton" architecture or more commonly as the ``von Neumann" model --- suffer from a well-known bottleneck. This bottleneck is due to the serial data transfer between separate memory and computing units. Thus, there has been increasing interest in the exploration of unconventional computing paradigms, ranging from in-memory computing, stochastic computing, neuromorphic computing, and reservoir computing, among others. Optical computing has also been explored as a means to leverage particular advantages of working with light over RF signals, or simply as a basis for the implementation of some of the above-mentioned paradigms.

Photonics can make a difference in computing, most notably in terms of:

* Bandwidth: photonics operate on lightwaves (100s of THz) while digital electronics operate on RF waves (GHz). The difference in potential bandwidth is 5 orders of magnitude. Massive parallelism can be achieved through frequency-multiplexing, combs can generate on the order of lines. Data in each line can be processed in parallel.
* Near-dissipationless dynamics and low-loss transmission: photons can propagate with negligible losses in free-space setups and very low losses in some integrated platforms (e.g. losses demonstrated on lithium niobate on insulator and silicon-nitride platforms ~ 0.2 dB/cm in the C-band. It is anticipated that optical interconnects will also play a role in electronic intra-chip communication.

Thus, there is great promise in utilizing photonic platforms for computing applications, especially those that would benefit from fundamentally different paradigms that promise boosts in speed and efficiency, e.g. application-specific signal processing and AI, and which can exploit and benefit from the above-mentioned strengths of working with light.

1. Background on Photonics

This section presents a brief overview of photonics, with a focus on photonic integrated circuits (PICs). Much research has gone into finding suitable materials for integrated photonic devices and circuits. For most platforms, photolithography is used to pattern wafers for etching and material deposition. Unlike with electronic circuits, where electrons exist readily inside materials, photonic circuits require the means of light *generation*, usually achieved by generating electron-hole pairs within a gain medium through an external excitation process, and *detection*, through the combination of those electron-hole pairs. From a historical perspective, device-level design preceded circuit-level thinking for decades. This resulted in the specialization of platforms for these different devices, for example: indium phosphide (InP) for lasers, gallium arsenide (GaAs) for detectors, and lithium niobate LiNbO3 for high-speed modulators. Each of these devices were made under different fabrication processes which are fundamentally incompatible with the ones used for the other devices.

Optical simulation of devices and systems requires consideration of the relevant physics. The physics of light can be described by different theories, which follow the evolution of our understanding of the nature of light. In chronological order, these theories are ray optics, scalar wave optics, beam optics, Fourier optics, electromagnetic optics, and quantum electrodynamics. In photonic circuits, depending on whether device level or system-level simulation is desired, and also on the kinds of phenomena taken into account, different theories of light may be employed for the modeling process. For integrated photonic system-level implementations, the modeling process is mainly reliant on guided-wave optics (e.g. coupled-mode theory). The circuit-level simulations carried out in this dissertation are done using the S-matrix method. A simple example is shown below to show how it works.

Consider a laser source, described as a coherent, monochromatic light field. The electric field of this source of frequency, assuming a plane wave (i.e no spatial dependence), can be described by its amplitude *A* and initial phase as

This light can be coupled onto an integrated photonic chip by various means, most commonly through a grating coupler. The field after being coupled into the grating coupler is thus

where is the coupling coefficient of the GC which depends on the laser wavelength and the angle between the fiber and the normal. After being coupled into the on-chip waveguide, it propagates for some distance yielding:

where is the length of the waveguide [m], is the waveguide loss [], and [] is the propagation constant of the guided mode with effective refractive index. The effective index describes the approximate index encountered by the propagating field due to the fact that not all the mode travels inside the waveguide core. This means that some of the mode will leak into the surrounding medium which in turn causes the field to encounter an `effective' index that takes into account this effect. Other circuit building blocks can be described in a similar approach, whether they are active or passive elements, and even if they have nonlinearities. Our simulations utilize these time-domain representations due to the temporal nature of time-delay reservoir computing.

1. Background on Reservoir Computing

Reservoir computing (RC) is one type of supervised machine learning approaches. Initially known as “Echo-state networks”, RC were used in software implementations in the early 2000s to circumvent the problems of training large recurrent neural networks. Physical systems can be harnessed to implement RC, most notably in electronics and photonics. The ESN is a nonlinear dynamical system which exhibits particular characteristics, namely:

* the echo-state property (ESP); the system allows reverberations of past input that decay with time (fading memory),
* dimensionality expansion, usually achieved through a nonlinear activation function.

The ESP is a condition that necessitates the forgetting of past inputs such that the reservoir does not depend on its initial conditions. This is an essential feature of RC as it would allow the reservoir to respond to the same input in the same manner. This consistent mapping is critical such that a model trained on the output of the reservoir is able to generalize. Dimensionality expansion enables the separation of multiple classes with simple linear classifiers, even for tasks that would normally require larger and deeper neural networks. The idea is, by projecting the data onto a higher dimensional space, there would exist more hyperplanes that can separate well the various classes of data, which enables high performance accuracy. Any RC scheme consists of the following 3 layers:

* Input layer: where data is preprocessed (optional) and then injected into the reservoir,
* Reservoir layer: which consists of nonlinear nodes connected to each other,
* Output layer: where the reservoir states are collected by means of a readout mechanism.

Each layer is connected to the next one via a series of weighted connections. What separates reservoir computing from similar temporal neural networks, such as recurrent neural networks (RNNs), is that the weights between the input and reservoir layer can be set and fixed, as well as the interconnections between the nonlinear nodes, whereas in RNNs all the connections need to be optimized. Thus, in RC only the output layer is trained, which means that only least-squares methods are required to find the optimal weights.

RC can be realized using spatial or time-multiplexed networks. The latter concerns the use of a single nonlinear node and sampling it in time to obtain the desired node states. This work focuses on designing time delay reservoir computers using photonic circuits. This topic has been explored since 2011, and a detailed state of the art chapter addresses the most notable innovations in this field.

**Part II: Scientific Contributions**

1. Minimum complexity photonic reservoir computing

The first contribution builds on the existing literature and the contemporary theme of reservoir computing in photonic integrated circuits, rather than in bulk. This study shows how a simple device (shown in Fig.1), consisting of passive waveguides and a heater element for reconfigurability, can already perform as well as more complex schemes. The nonlinearity is implemented using the intensity-conversion nonlinearity provided at the photodetector. The performance was measured on a variety of standard benchmark tasks.

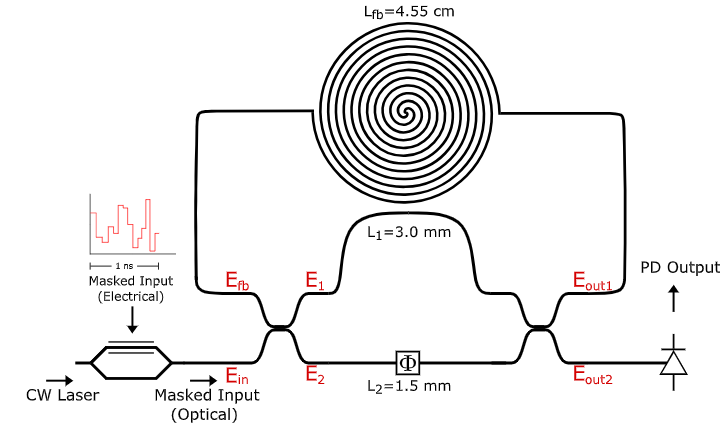


Fig. 1: Minimum complexity reconfigurable photonic reservoir computing architecture

The principle of operation is that the heating element changes the optical path length in the bottom waveguide, which causes a change in the output dynamics. Once the desirable dynamics are found, the circuit can be trained to solve benchmark tasks with high accuracy, such as the bitwise XOR, Mackey-Glass and Santa-Fe tasks.

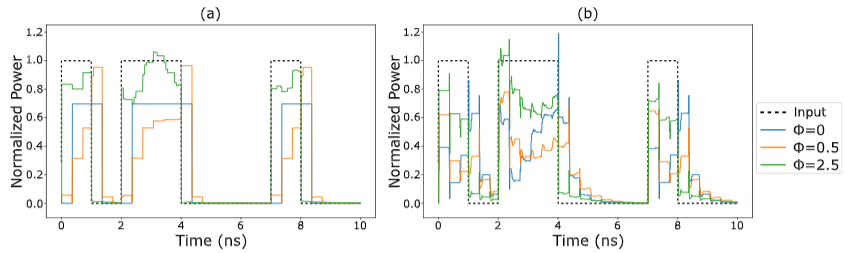


Fig. 2: System output for varying amounts of phase shift

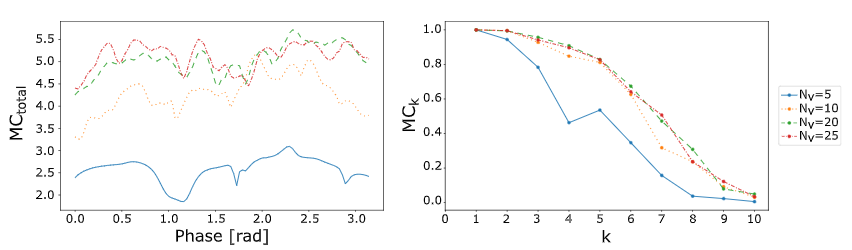


Fig. 3: Total memory capacity and peak individual capacities

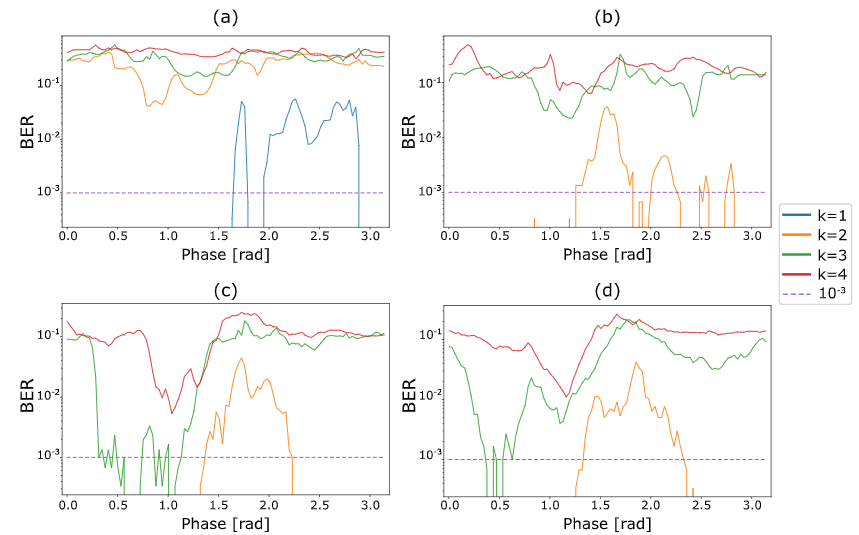


Fig. 4: XOR performance for various memory depths, under different values of applied phase shift.

Another contribution is that the feedback is integrated onto the chip, thanks to the the consideration of the low-loss LNOI platform. This study ultimately shows what a simple time-delay reservoir computing scheme can do, and serves as a benchmark to compare against when more complex schemes are considered.

1. Mask-less asynchronous photonic reservoir computing

One of the key limitations in time-delay reservoir computing is that the time-multiplexing of nodes limits the effective processing speed of the system. This is further augmented by the widespread use of an input mask, which further reduces the effective processing speed. Additionally, it adds constraints in terms of complexity and energy consumption since a domain conversion, from an incoming optical signal to electronic and back to optical, may be required to apply the mask. In this contribution, we explore the possibility of using the reservoir without the input mask by engineering the feedback length to keep the system from falling to steady state for the duration of the input sample, which is the fundamental role of the input mask. In doing so, the input mask can be bypassed and the reservoir can solve a wide variety of tasks at much faster speeds.

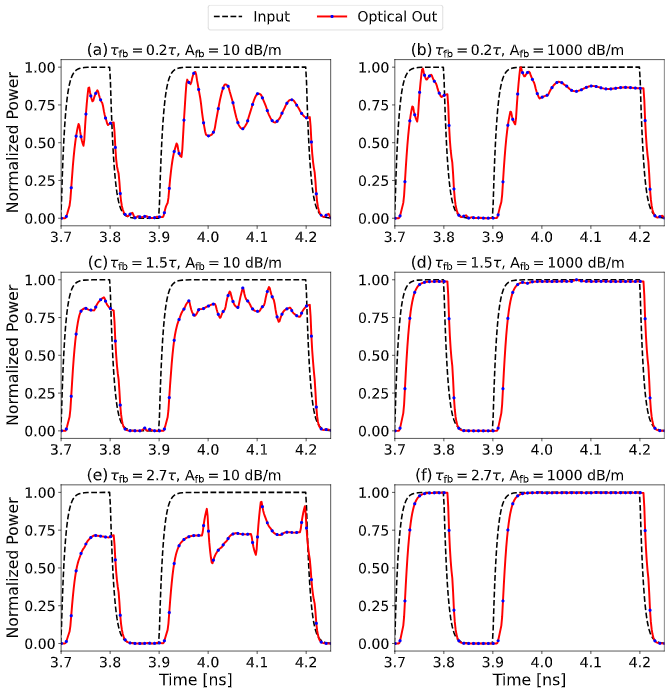


Fig. 5: Shorter feedback lengths allow the signal to interact with itself multiple times, allowing an emulation of the input masking effect. This is possible when considering low-loss platforms.

We test the reservoir on a range of task-independent metrics, as well as some standard benchmark tasks. The numerical results show good performance on a wide variety of benchmark tasks, which is comparable to other schemes in the literature.

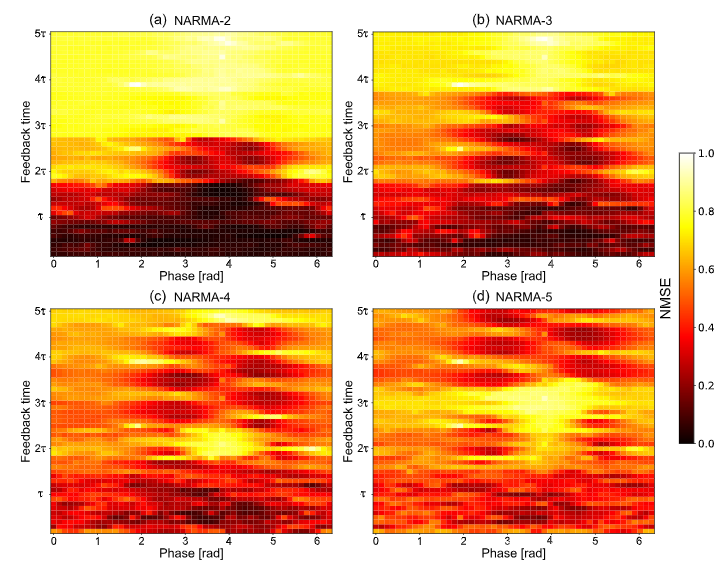


Fig. 6: The normalized mean square error (NMSE) is shown for a variety of tasks under different feedback times and phase shifts.

1. Effect of bandwidth-limited readout

We also consider the effects of the electronic readout on the performance of this architecture. The readout has a pivotal role, being the intermediary between the optical reservoir and the acquired signal. The readout, consisting of a photodiode and associated circuitry, imprints imperfections onto the signal which would be used for training the least-squares model. To this end, we consider a simple photodiode and noise model, and take into account the subsequent analog-to-digital conversion at the digital oscilloscope. For the ADC, we also shed light on the performance using a single acquisition at lower effective number of bits to verify that the scheme can work under these conditions for some tasks. The ENOB considered are typically found on high-end oscilloscopes.



Table 1: The scheme is able to solve some tasks even when considering a single acquisition routine with a lower effective number of bits.

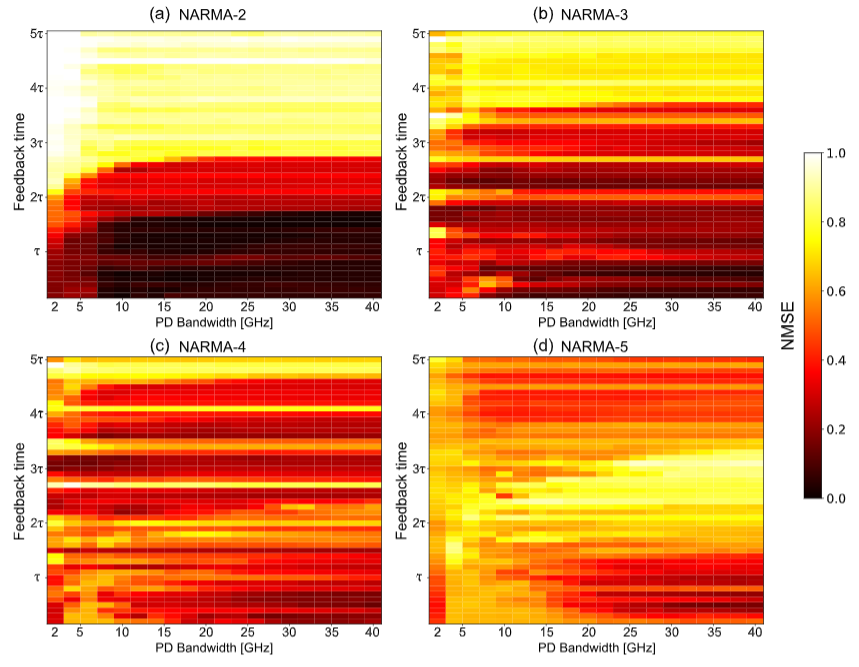


Fig. 7: The performance is recorded for various photodetector bandwidths and feedback times, showing high tolerance for lower bandwidths for certain tasks.

1. Experimental demonstration

In this last study, we report the experimental demonstration of the minimum complexity reservoir computing scheme under the mask-less condition on the Santa-Fe benchmark task, achieving excellent results. The chip was fabricated by the fabrication team at RMIT’s Integrated Photonics and Applications Center. The chip layout was done using IPKISS and InPAC’s process development kit (pdk). Design rule checking was done manually to ensure:

* minimal amount of bends overlap with the electron beam lithography (EBL) grid lines;
* minimum distance between nearby waveguides is respected;
* DC pad spacing (200 m) corresponds to the pitch of the DC probe;
* RF pad spacing corresponds to the pitch or the RF probe.

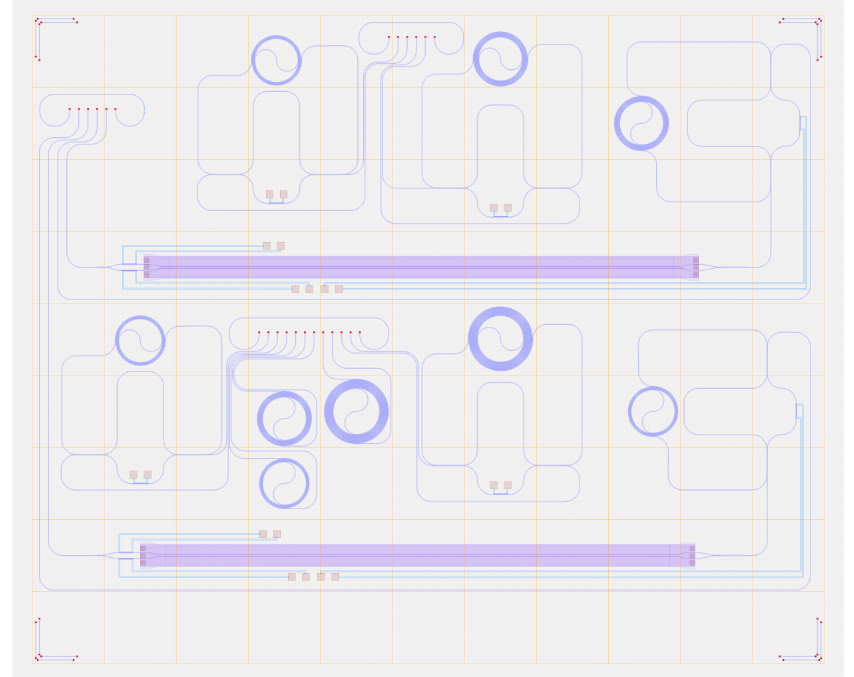


Fig. 8: Chip layout generated using IPKISS. Multiple test devices are considered with varying feedback lengths, using on-chip modulators, as well as test structures.

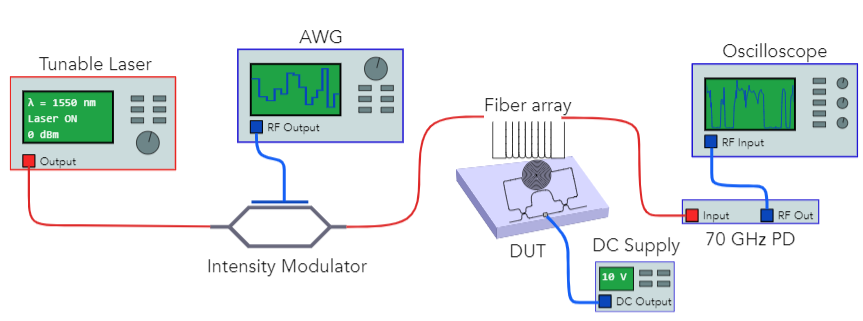


Fig. 9: Experimental setup

The experimental results on the Santa Fe task are shown in Fig. 10 across all the datasets gathered. The datasets represent the sweep done on the wavelength and heater voltage. Two devices were considered, the one using 2.84 cm feedback loop (R284) and the one using 1.88 cm feedback loop (R188). Normally the Santa Fe task is considered for the one-step ahead prediction, but we have also documented the results for 2 steps and 3 steps prediction. It can be seen that R188 performs better than R284, due to the lower feedback losses encountered.

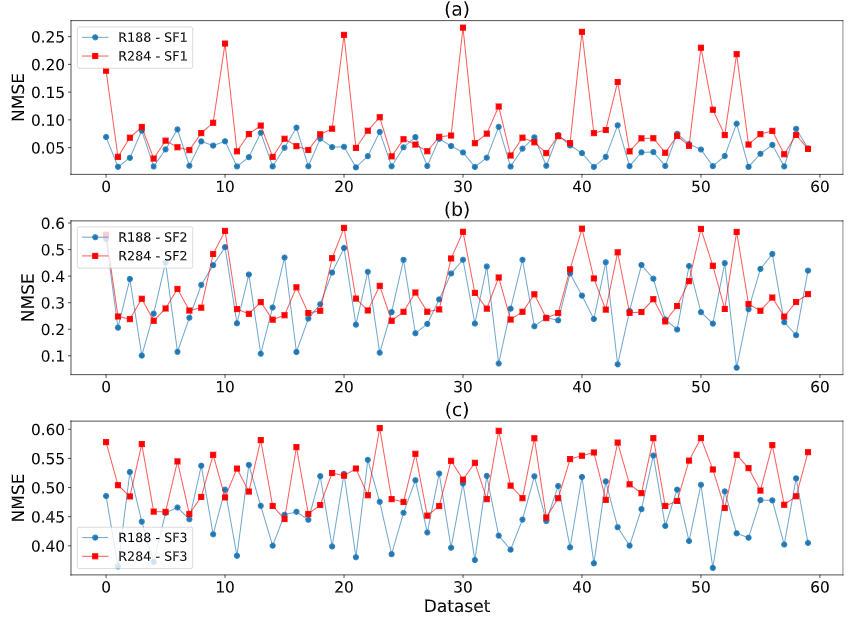


Fig. 10: NMSE obtained with the various datasets for the Santa Fe task for (a) 1 step ahead prediction, (b) 2 steps ahead prediction, and (c) 3 steps ahead prediction.

1. Conclusions

The main theme explored throughout this work is the minimization of complexity of photonic RC schemes, while still obtaining good results. To that aim, we have launched multiple investigations. The first concerns proposing a novel integrated photonic RC architecture that is reconfigurable using one element, in the form of a phase shifter, which controls two parameters simultaneously; the feedback strength and feedback phase. Tuning the feedback strength in this manner does not require power to be coupled out of the system, as is typically achieved by using an optical attenuator. The integration of the feedback waveguide is enabled by the low-loss LNOI platform or any other low-loss platform, such as silicon nitride. Our simplified design enabled good performance on some RC benchmark tasks, while at the same time significantly reducing the hardware complexity of the reservoir and auxiliary equipment. The second investigation concerns further reduction of complexity by disposing with the input mask at the input layer and engineering the feedback length to yield dynamics which stay in transience during the input perturbation, which is essentially the aim of the input mask. The reservoir's performance in the asynchronous regime was evaluated using a variety of task-independent metrics and benchmark tasks. The results showed comparable performance to the literature for the studied tasks. In summary, the results presented in this dissertation suggest that such minimal RC schemes can achieve on-par performance with schemes that either require costly auxiliary instruments, or schemes that use a multitude of tunable parameters. Finally, the effects of the bandwidth-limited, noisy readout were explored, showing how it affects the performance for various tasks. It was concluded that the nature of the task determines the robustness of the system's performance to lower bandwidths. Finally, the experimental demonstration was presented and the scheme was shown to solve the Santa Fe benchmark task with excellent results.

There are several key limitations of the proposed architecture, particularly at the proposed high speeds. The first is the need for high speed readout equipment to achieve higher bandwidths. This means high bandwidth photodetectors and oscilloscopes, which are costly and bulky. Like other chip-based schemes, it is limited to solving tasks with moderate memory and below. For example, solving the NARMA-10 task with this scheme, especially when considering readout noise, becomes very difficult. Furthermore, while we brushed on the topic of solving multiple tasks at the same time, further investigation is still needed. In particular, it would be interesting to investigate the use of multiple wavelengths, as could be done using an integrated micro-comb source, which is possible on the LNOI platform. Furthermore, the rich variety of optical nonlinearities accessible on LNOI can be explored and its RC performance compared against this minimalist scheme. Finally, and this is not specific to this scheme: photonic systems suffer inherently from wavelength drift due to refractive index change in response to changing temperatures. While this is possible to compensate in a lab setting, it is not straightforward in a packaged environment for real-world applications. While several studies have explored means to mitigate this problem, there is still room for improvement. One way to handle this using the proposed architecture could be through using either an FPGA or a co-packaged ASIC that implements a zero-calibration point algorithm to compensate for temperature drift. The ASIC would take as the input a portion of the detected output power, and accordingly adjust the phase shifter voltage to perform the compensation.