

Fig. 1. Three different approaches to analog electronic and optical computing that arrays performance computing that arrays performance crossbar arrays performance crossbar arrays performance crossbar arrays performance c computing in-memory by applying inputs as voltages to rows of the array, and storing weights in flash, memristors, or phase change memory as the conductance between two points  $[2]$ ,  $[3]$ , produce  $\mathbb{P}_\mathcal{A}$  is proportional to the multiplication. (b)  $\mathbb{P}^\bullet$  is proportional to the multiplication integrates charge at a coherent detector at a coherent detector and detector and detector and detector and accumulates Macs over time stepstification Broadcast and wavelength uses modulation multiplication multiplication multiplication multiplication multiplication multiplication multiplication multiplication multiplication mu  $\frac{1}{2}$ ,  $\frac{3}{2}$ ,

multipliers for deep learning do not dynamically vary precisions  $\frac{1}{2}$ to account for the precision sensitivity of different layers. At  $\mathbf{r}_\mathbf{y}$  $\mathbf{u} = -\frac{1}{2}$  and  $\mathbf{u} = -\frac{1}{2}$  and  $\mathbf{v} = -\frac{1}{2}$  and  $\mathbf{v} = -\frac{1}{2}$ determined at determined at design time for each network  $\mathcal{L}$  $\left(24\right)$ , or utilize mixed precision by using a digital processor for  $\frac{12}{3}$ precision-sensitive operations, such as backpropagation or the  $\frac{1}{2}$ ,  $\frac{1}{2}$  $\frac{1}{\sqrt{2}}$  in this work, we propose extending analog computing analog computing archi-

 $\mathcal{L}_{\text{c}}$  to support dynamic precision that can be selected by a best  $\mathcal{L}_{\text{c}}$ programmer or compiler, analogous to bit precision in digital  $\mathcal{M}_{\rm tot}$  and it is possible to trade of  $\mathcal{M}_{\rm tot}$  is possible to trade of  $\mathcal{M}_{\rm tot}$ performance metrics, such as energy efficiency, through  $\mathcal{P}_\text{max}$ area, to improve the precision of the analog computing engine. The analog computing engine. The analog computing engine. The second  $B_{\rm eff}$  repeating the same operation multiple times and averaging times and averaging and averaging times and averaging the results (as demonstrated by multi-memorial by multi-memorial synapses  $(24)$ ), and  $(24)$ precision can be improved at the cost of expending more energy.  $\mathcal{L} = \{ \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1} \}$  , we discuss how redundant coding (repeating  $\mathbf{v}_1$  $\frac{1}{2}$   $\frac{y_1^2}{y_1^2}$  in space or time can be applied to both can be applied analog electronic and optical computing architectures to enable architectures to enable architectures to enabl  $\mathbf{0.25}_{\textbf{0.00}}\times\mathbf{0.10}_{\textbf{0.00}}$ 

A key challenge for deploying  $\mathcal{A}$  and depend networks with dynamic  $\mathcal{A}$ precision is determining the optimal precision of different layers of the neural network given a hardware performance target. In the second state  $\mathcal{P}_{\text{max}}$ Section V, we take the tackle this by solving and  $\mathbb{P}^{\mathcal{P}}$  and problem. Problem  $W = \frac{1}{2} \int_{0}^{\infty} \frac{1}{2} \int_{0}^{\$ multiply-accumulate (MAC) of redundant coding and the result-of result-of redundant coding and the result-of  $\mathbf{w}$ ing precision. We define a constrained optimization. We define a constrained optimization problem to the constraints of the constrained optimization problem to the constraints of the constraints of the constraints of the maximize the original objective of the neural network subjective of the neural network subject to  $\mathcal{N}$ and constraint on total energy consumed, where the energy consumed, where the energy  $\mathbf{M}^{\prime}$ 





Fig. 3. Dynamic precision with redundant coding; redundant coding; resistive example. Changes to the architecture arrays to the arrays to the architecture are shown in red. We are shown in red. We are shown in red. We are use K to denote the number of times and repeated for times are repeated for K clock cycles, in (b) the same inputs are repeated for K clock cycles, in (b) the same inputs and weights are repeated for K clock cycles, in (b repeated, and in (c) only certain rows of **W** are repeated.



$$
\mathbf{W} = \begin{bmatrix} \mathbf{W} & \mathbf{W} & \dots & \mathbf{W} \end{bmatrix} \begin{bmatrix} \mathbf{W} & \mathbf{W} & \mathbf{W} & \mathbf{W} \end{bmatrix}
$$





Authorized licensed use limited to: Queen's University. Downloaded on November 24,2022 at 14:38:22 UTC from IEEE Xplore. Restrictions apply.

 $\mathbf{T}=\mathbf{E}$  III D





- $\mathbf{x} = \mathbf{y} \mathbf{y}$  . And  $\mathbf{y} = \mathbf{y}$  and  $\mathbf{y} = \mathbf{y}$ neural networks: A survey," *IEEE Access*, vol. 7, pp. 175827–175841, 2019.
- $\mathcal{L} = \frac{M}{\pi} \mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3$  intelligence and neuromorphics for artificial intelligence and neuromorphic and neuromorphic and neuromorphic and neuromorphic and neuromorphic and neuromorphic and neuromor computing," *Nature Photon.*, vol. 15, no. 2, pp. 102–114, Feb. 2021, doi: [10.1038/s41566-020-00754-y.](https://dx.doi.org/10.1038/s41566-020-00754-y)
- $\mathcal{I} = \mathcal{I} \cup \mathcal{I$ *Photon.*, vol. 11, no. 7, pp. 441–446, 2017.
- [8] D. A. B. Miller, "Perfect optics with imperfect components," *Optica* vol. 2, no. 8, pp. 747–750, Aug. 2015. [Online]. Aug. 2015. [Online]. Available: [http://www.](http://www.osapublishing.org/optica/abstract.cfm{?}URI=optica-2-8-747)
- osapublishing.org/optical/abstract.cfm?uRi=optica-2-8-7474.cfm?uRi=optica-2-8-74747.cfm?uRi=optica-2-8-7474747  $\zeta$  ,  $\zeta$  M. Misculio and V. Sorger, "Photonic tensor cores for machine learning,"  $\zeta$  machine learning,  $\zeta$ *Appl. Phys. Rev.*, vol. 7, 2020, Art. no. 031404.
- $\mathcal{A}_1=\mathcal{A}_2=\mathcal{A}_1+\mathcal{A}_2=\mathcal{A}_2=\mathcal{A}_1+\mathcal{A}_2$ i and computer,  $\mathcal{A}_2=\mathcal{A}_1+\mathcal{A}_2$ able:<https://openai.com/blog/ai-and-compute/> $\mathcal{P}_\text{max}$
- $\mathbf{1}_{11}$  between  $\mathcal{T}_{11}$  B. Jacob et al.,  $\mathcal{T}_{11}$  between and training of neural networks for efficient  $\mathcal{T}_{12}$ integer-arithmetic-only inference," in *Proc. IEEE/CVF Conf. Comput. Vis.*
- *Pattern Recognit.*, 2018, pp. 2704–2713. [12] R. Krishnamoorthi, "Quantizing deep convolutional networks for efficient inference: A whitepaper," 2018, *arXiv:1806.08342*.
- $\mathcal{H}_1$  ,  $\mathcal{H}_2$  ,  $\mathcal{H}_3$  ,  $\mathcal{H}_4$  ,  $\mathcal{H}_5$  training  $\mathcal{H}_4$  bit  $\mathcal{H}_5$ of convolutional networks for rapid-deployment," in *Proc. Adv. Neural*
- *Inf. Process. Syst.*, H. Wallach et al., Eds., vol. 32, 2019, pp. 7950–7958.  $\frac{1}{\sqrt{2}}$  (original bless). As a single sin  $\mathcal{A}_1$  ,  $\mathcal{A} = \mathcal{A}_1$  ,  $\mathcal{A}_2$  and  $\mathcal{A}_3$  are  $\mathcal{A}_4$  in  $\mathcal{A}_5$  . And  $\mathcal{A}_6$  is the  $\mathcal{A}_7$  subset of  $\mathcal{A}_8$  ,  $\mathcal{A}_7$  ,  $\mathcal{A}_8$  ,  $\mathcal{A}_9$  ,  $\mathcal{A}_9$  ,  $\mathcal{A}_9$  ,  $\mathcal{A}_9$  ,  $\mathcal{A}_9$  ,
- Hessian are distinct and networks  $\mathcal{F}_{\alpha}$  in a set  $\mathcal{F}_{\alpha}$  with mixed-precision,  $\mathcal{F}_{\alpha}$  $V'$  is the computation of  $V'$ <sub>rap</sub>  $V'$ <sub>rap</sub>  $\rightarrow$  2019, pp. 2019, pp.
- $\mathbf{15}$   $\mathbf{15}$   $\mathbf{15}$   $\mathbf{15}$   $\mathbf{16}$   $\mathbf{17}$   $\mathbf{18}$  and  $\mathbf{17}$  are trace-weighted quantization and  $\mathbf{17}$ of neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 18518–18 $V_{\rm 4}$  , 2020. 2020. 2020. 2020. 2020. 2020. 2020. 2020. 2020. 2020
- $\mathbf{1}_{16}$  K.  $\mathbf{1}_{16}$  K. Hang,  $\mathbf{1}_{16}$   $\mathbf{$ automated quantization with  $\mathbf{v}' = \mathbf{v}'$  in  $\mathbf{v}' = \mathbf{v}'$ *Comput. Vis. Pattern Recognit.*, 2019, pp. 8604–8612.
- $\mathbf{1}_{11}=\mathbf{1}_{11}=\mathbf{1}_{12}=\mathbf{1}_{13}=\mathbf{1}_{14}=\mathbf{1}_{15}=\mathbf{1}_{16}=\mathbf{1}_{17}=\mathbf{1}_{18}=\mathbf{1}_{18}=\mathbf{1}_{18}=\mathbf{1}_{18}=\mathbf{1}_{18}=\mathbf{1}_{18}=\mathbf{1}_{18}=\mathbf{1}_{18}=\mathbf{1}_{18}=\mathbf{1}_{18}=\mathbf{1}_{18}=\mathbf{1}_{18}=\mathbf{1}_{18}=\mathbf{1}_{18}=\mathbf{1}_{18}=\mathbf{1}_{18}=\mathbf{$ neural architecture search," 2018, *arXiv:1812.00090*.
- $\mathbf{1} = \mathbf{1} \cdot \mathbf{1} \cdot \mathbf{1} = \mathbf{2} \cdot \mathbf{1} \cdot \mathbf{1} = \mathbf{1} \cdot \mathbf{1} \cdot \mathbf{1} = \mathbf{1}$  $\mathbf{v}'$ pp. 13166–13175.
- $\mathbf{1}_{\mathcal{N}}$  i. Hubara, I. Hanani, I. Hanani, R. Banner,  $\mathbf{M}$  i.  $\mathbf{M}$  is  $\mathbf{M}$  in  $\mathbf{M}$  in  $\mathbf{M}$ ing post training neural quantization: Layer-wise calibration:  $L$ ayer-wise calibration and integers  $\sim$ programming," vol. abs/2006.10518, 2020.
- $\mathcal{Z}_{20}$  ,  $\mathcal{Z}_{20}$  and  $\mathcal{Z}_{20}$  and  $\mathcal{Z}_{20}$  and  $\mathcal{Z}_{20}$   $\mathcal{Z}_{20}$  point  $\mathcal{Z}_{20}$  and define  $\mathcal{Z}_{20}$ convolutional networks," in *Proc. 33rd Int. Conf. Mach. Learn.*, M. F.  $B_{\mathbf{A}}$  ,  $\mathbf{A}$  ,  $\mathcal{P}_\text{2D}$  and  $\mathcal{P}_\text{2D}$  are as  $\mathcal{P}_\text{2D}$  . And the set of the s  $\gamma$  and  $\gamma$  is  $\gamma$
- $\mathcal{L}_1$  S. Uhlich et al., "Mixed precision DNNs: All you need is a good precision DNNs: All you need is a good parametrization," in *Proc. Int. Conf. Learn. Representations*, 2020,  $\frac{1}{2}$ p $\sim$  67–68.
- $\mathcal{L}_{12}$  NVIDIA  $\mathcal{L}_{22}$  are architecture,  $\mathcal{L}_{12}$  are p.  $\mathcal{L}_{23}$  are p.  $\mathcal{L}_{24}$  are p.  $\mathcal{L}_{25}$  $20\%$  and  $20\%$  are  $20\%$  . Available: https://www.nvidia.com/content/dam/en-za/ [Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf](https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf)
- $\mathcal{L}_{\mathcal{A}}$  ,  $\mathcal{L}_{\mathcal{B}}$  ,  $\mathcal{L}_{\mathcal{A}}$  , and D. Englund,  $\mathcal{L}_{\mathcal{B}}$  , and  $\mathcal{L}_{\mathcal{B}}$  ,  $\mathcal{L}_{\mathcal{B}}$  , and  $\mathcal{L}_{\mathcal{$ scale optical neural networks based on photoelectric multiplication," *Phys. Rev. X*, vol. 9, May 2019, Art. no. 021032.
- [24] I. Boybat et al., "Neuromorphic computing with multi-memristive synapses," *Nature Commun.*, vol. 9, no. 1, pp. 1–12, 2018. [25] M. Gallo et al., "Mixed-precision in-memory computing," *Nature Elec-*
- *tron.*, vol. 1, pp. 246–253, 2018.  $\mathcal{Z}_{\mathcal{C}}$  S. R. Nandakumar et al., "Mixed-precision architecture based on computer based o
- tational memory for training deep neural networks," in *Proc. IEEE Int.*
- *System System Syst*  $\mathbb{P}_2$  . E. E. e.f.,  $\mathbb{P}_2 = \frac{1}{2}$  ,  $\mathbb{P}_2 = \frac{1}{2}$  , and in expectation based on in-memory  $\mathbb{P}_2$ computing," *IBM J. Res. Dev.*, vol. 63, no. 6, pp. 7:1–7:16, 2019.
- $\mathcal{Z}_2$ 8] S. M. Nandakumar et al., "Mixed-precision deep learning based on computational memory," *Front. Neurosci.*, vol. 14, 2020, Art. no. 406. [Online]. Available:<https://www.frontiersin.org/article/10.3389/fnins.2020.00406>9/fnins.2020
- $\begin{bmatrix} 2 & 3 \end{bmatrix}$  G. G. G. W. Burr et al.,  $\begin{bmatrix} 2 & 3 \end{bmatrix}$  with  $\begin{bmatrix} 2 & 3 \end{bmatrix}$  with  $\begin{bmatrix} 2 & 1 \end{bmatrix}$ memory," *Adv. Phys.: X*, vol. 2, no. 1, pp. 89–124, 2017,
- doi: [10.1080/23746149.2016.1259585.](https://dx.doi.org/10.1080/23746149.2016.1259585)  $\mathbf{x}_t = \mathbf{y}_t + \mathbf{y}_t^T \mathbf{y}_t^T \mathbf{y}_t^T$  and  $\mathbf{y}_t^T$  and  $\mathbf{y}_t^T$  and  $\mathbf{y}_t^T$  and computed devices for computational devices for computations for computations for computations for computations of computatio  $\mathbf{v}'$   $\mathbf{v}'$   $\mathbf{v}'$   $\mathbf{v}'$   $\mathbf{v}'$   $\mathbf{v}'$   $\mathbf{v}'$
- $\mathbf{1}_{\mathcal{M}}$  H.-H.-T. Ambrogio, P. Narayanan, R. Shelby, and G. Burr,  $\mathcal{R}$  , and G. Burr, "Recent B progress in analog memory-based accelerators for deep learning," *J. Phys. D: Appl. Phys.*, vol. 51, no. 28, 2018, Art. no. 283001.
- $\mathbf{x}_1 = -1$  et al.,  $\mathbf{z}_2 = -1$  et al.,  $\mathbf{z}_3 = -1$  et al.,  $\mathbf{z}_4 = -1$  et al.,  $\mathbf{z}_5 = -1$  et al.,  $\mathbf{z}_6 = -1$ arrays," *Nature Mach. Intell.*, vol. 1, no. 1, pp. 49–57, 2019.  $\overline{33}$  and  $\overline{33}$  arrays,  $\overline{33}$  and  $\overline{33}$  arrays,  $\overline{33}$  and  $\overline{33}$  arrays,  $\overline{33}$  arrays,  $\overline{33}$
- *Nature Electron.*, vol. 2, no. 3, pp. 115–124, 2019.  $\mathbf{x}_k = \mathbf{x}_k$ grad et  $\mathbf{E} = \left[\mathbf{E}_{k+1} \mathbf{E}_{k+2} \right]$ ing advantages of resistive memory crossbarges of resistive memory crossbarges of  $\mathbf{E}$
- based computation and its application to sparse coding," *Front. Neurosci.* vol. 9, 2016, Art. no. 484. [Online]. Available: [https://www.frontiersin.](https://www.frontiersin.org/article/10.3389/fnins.2015.00484) [org/article/10.3389/fnins.2015.00484](https://www.frontiersin.org/article/10.3389/fnins.2015.00484)
- $\mathcal{A}$  M. R. Mahmoodi, M.  $\mathcal{B}$ . M.  $\mathcal{B}$ . Strukov, "Versatile stochastic stochas dot product circuits based on nonvolatile memories for high performance neurocomputing and neurooptimization," *Nature Commun.*, vol. 10, no. 1,  $N_{\rm eff}$  2019,  $N_{\rm eff}$  [10.1038/s41467-019-13103-7.](https://dx.doi.org/10.1038/s41467-019-13103-7)1038/s41467-019-13103-7.1038/s4146
- $\mathbb{R}^3$  a. N. Tait, M. A. Nahmi $\mathbb{R}^3$  and  $\mathbb{R}^4$  and  $\mathbb{R}^4$  and  $\mathbb{R}^4$  and  $\mathbb{R}^4$  and  $\mathbb{R}^4$ weight: An integrated network for scalable photonic spike processing," *J.*
- $\sum_{i=1}^{n} \frac{V^{i}}{i!} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{i!} \sum_{j=1}^{n} \frac{1}{j!} \sum_{j=1}^{n} \$  $\mathbf{y} = \mathbf{T} \cdot \mathbf{F} \mathbf{A}$  and the limit of alleged with neuromorphic photonics,  $\mathbf{y} = \mathbf{y} \cdot \mathbf{A}$ *J. Lightw. Technol.*, vol. 37, no. 5, pp. 1515–1534, Mar. 2019. [Online].
- $\mathcal{A}$  and  $\mathcal{A}$  is the symplectic space of  $\mathcal{A}$  absorption  $\mathcal{A}$ [38] H. Peng, M. A. Nahmias, T. F. de Lima, A. N. Tait, and B. J. Shastri, "Neuromorphic photonic integrated circuits," *IEEE J. Sel. Topics Quantum*
- *Electron.*, vol. 24, no. 6, pp. 1–15, Nov./Dec. 2018.  $\mathcal{J}_\infty$  . The compatibility  $\mathcal{J}_\infty$  is and  $\mathcal{J}_\infty$  in  $\mathcal{J}_\infty$  . One computed the compatibility compatibility of  $\mathcal{J}_\infty$ optical signal processor," *Opt. Exp.*, vol. 20, no. 12, pp. 13560–13565, Jun. 2012. [Online]. Available: [http://www.opticsexpress.org/abstract.](http://www.opticsexpress.org/abstract.cfm{?}URI=oe-20-12-13560)
- $\overline{\mathcal{S}}=\mathcal{S}^{\prime}$  and  $\overline{\mathcal{S}}^{\prime}$  and  $\overline{\mathcal{S}}^{\prime}$  and  $\overline{\mathcal{S}}^{\prime}$  $\mathcal{A}_1$  ,  $\mathcal{A}_2$  bandari et al.,  $\mathcal{B}_2$  photonics and analog photonics for convocations for convocations for convo lutional networks  $(E_{A_1}, E_{B_2}, \ldots, E_{B_n})$
- *Electron.*, vol. 26, no. 1, pp. 1–13, Jan./Feb. 2020.  $\begin{bmatrix} 1 & A & A \\ 1 & A & B \end{bmatrix}$ ,  $\phi^2 = \frac{1}{2} \begin{bmatrix} (\mathbf{E}^2 \times \mathbf{E}^2)^{-1} & \mathbf{E}^2 \times \mathbf{E}^2 \\ \mathbf{E}^2 \times \mathbf{E}^2 & \mathbf{E}^2 \times \mathbf{E}^2 \end{bmatrix}$ <br>  $\mathbf{E} = \mathbf{E} \mathbf{E$
- *Appl.*, vol. 11, Jun. 2019, Art. no. 064043, doi: [10.1103/PhysRevAp](https://dx.doi.org/10.1103/PhysRevApplied.11.064043) $p_1$ .064 $p_2$ [42] A. N. Tait et al., "Microring weight banks," *IEEE J. Sel. Topics Quantum*
- *Electron.*, vol. 22, no. 6, pp. 312–325, Nov./Dec. 2016.  $\mathcal{A}_\mathcal{A}$  a. Tait et al., "Neuromorphic photonic networks using silicon photonic photonic
- weight banks," *Sci. Rep.*, vol. 7, no. 1, pp. 1–10, 2017.  $\frac{1}{4}$  M. A. H. A. I.,  $\frac{1}{4}$  and  $\frac{1}{4}$  in a photonic integrated in a photonic integrated in a photonic integrated in
- $\mathbf{C}$ circ $\mathbf{C}$  $\mathcal{A}$  M. Hayat,  $\mathcal{B}^H$  M. Hayat, and J. Gubner, "Shot-noise-limited performance pe
- of optical neural networks," *IEEE Trans. Neural Netw.*, vol. 7, no. 3,  $p_0 = p_0$ ,  $p_1 = 1$
- $\frac{1}{2} \sum_{i=1}^{n} \frac{1}{i} \sum_{j=1}^{n} \frac{1}{i} \sum_{i=1}^{n} \frac{1}{i} \sum_{i=1}^{n} \frac{1}{i} \frac{1}{i} \sum_{i=1}^{n} \frac{1}{i} \frac{1}{i} \sum_{i=1}^{n} \frac{1}{i} \sum_{j=1}^{n} \frac{1}{i} \sum_{i=1}^{n} \frac{1}{i} \sum_{i=1}^{n} \frac{1}{i} \sum_{i=1}^{n} \frac{1}{i} \sum_{i=1}^{n} \frac{1}{i} \sum_{i=1}^{n} \frac$ algorithm accelerator," in *Proc. Int. Joint Conf. Neural Netw.*, 2016, pp. 929–938.
- $\mathcal{A}_1$  M. M. Hu et al.,  $\mathcal{B}_2$  and  $\mathcal{B}_3$  are neuromorphic computing: Product engine for neuromorphic computing:  $\mathcal{B}_3$ ming 1T1M crossbar to accelerate matrix-vector multiplication," in *Proc. 53nd ACM/EDAC/IEEE Des. Automat. Conf.*, 2016, pp. 1–6.
- $\frac{1}{2}$  M. A. A. Mias,  $\frac{1}{2}$  and  $\frac{1}{2}$  u. U. S.  $\frac{1}{2}$  and  $\frac{1}{2}$  u. Patentic computing,  $\frac{1}{2}$ ر<br>10,656, Apr. 2020. 2020. 2020. 2020. 2020. 2020. 2020. 2020. 2020. 2020. 2020. 2020. 2020. 2020. 2020. 2020. 2
- $\mathcal{A}_{\mathcal{A}}$  b. Feinberg, Making memorial memori





**Sahaj Garg** received the B.S. degree in computer sci-