Devices and Algorithms for Analog Deep Learning

MURAT ONEN

monen@mit.edu

Digital Processors are Insufficient to Advance AI

System Performance (TOPS/W)

CPU	GPU	ASIC
0.01	1	10

Digital Limit 100

Performance Requirement: >1000

2025 Estimates

$1B

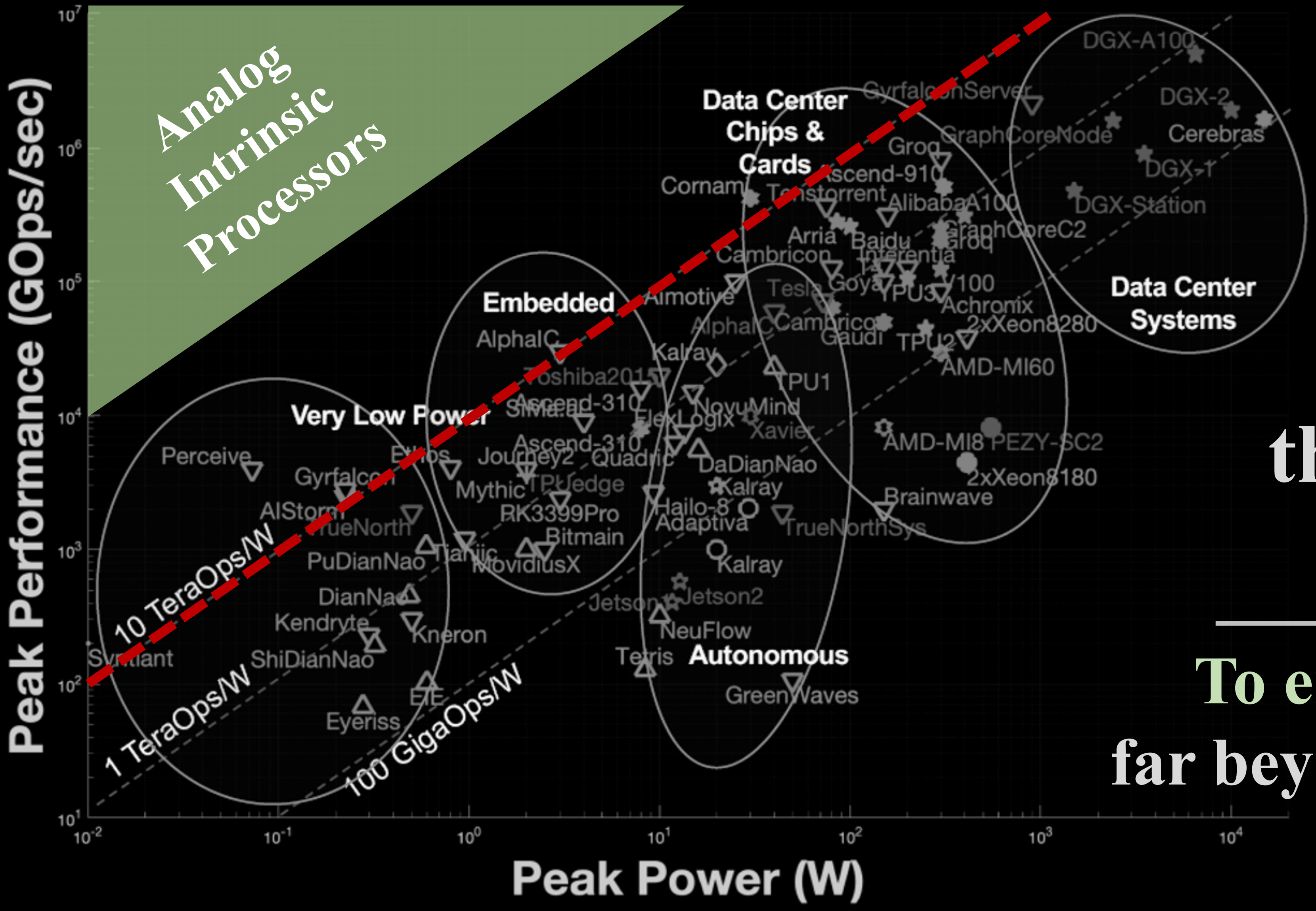
Training cost of one large model.

Higher power consumption than **NYC**

Era-Defining Problem

- ❖ Retail
- ❖ BFSI
- ❖ Manufacturing
- ❖ Marketing
- ❖ Healthcare
- ❖ Automotive
- ❖ Defense
- ❖ Telecom/IT

T. Brown, et al., Language Models are Few-shot Learners, *NeurIPS,* (2020).
E. Strubell, A. Ganesh, A. McCallum, Energy and Policy Considerations for Deep Learning in NLP, *ACL,* (2020).
N.C. Thompson, et al. Deep Learning's Diminishing Returns: The Cost of Improvement is Becoming Unsustainable, *IEEE Spectrum,* (2021)



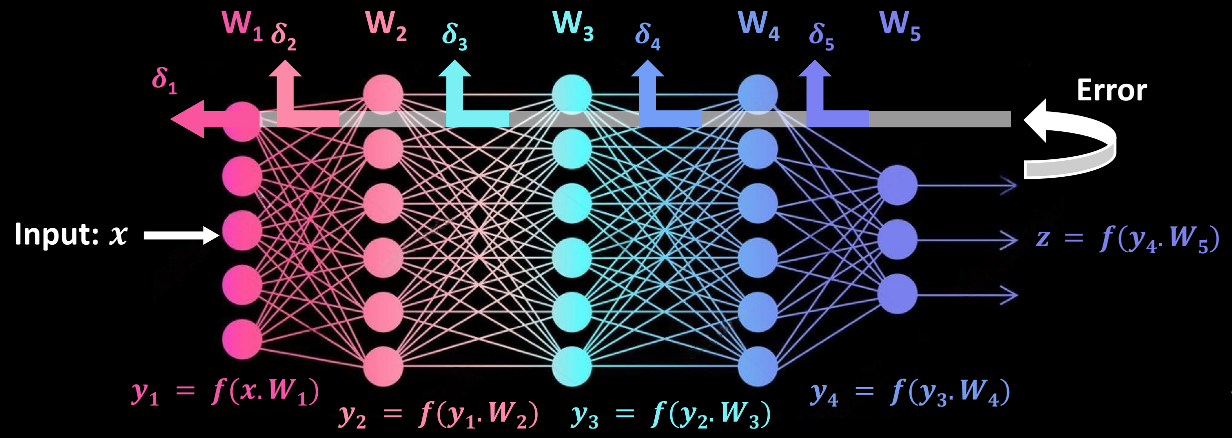
Promising **100× higher performance** than the leading digital ASICs.

To enable advanced AI far beyond current reach.

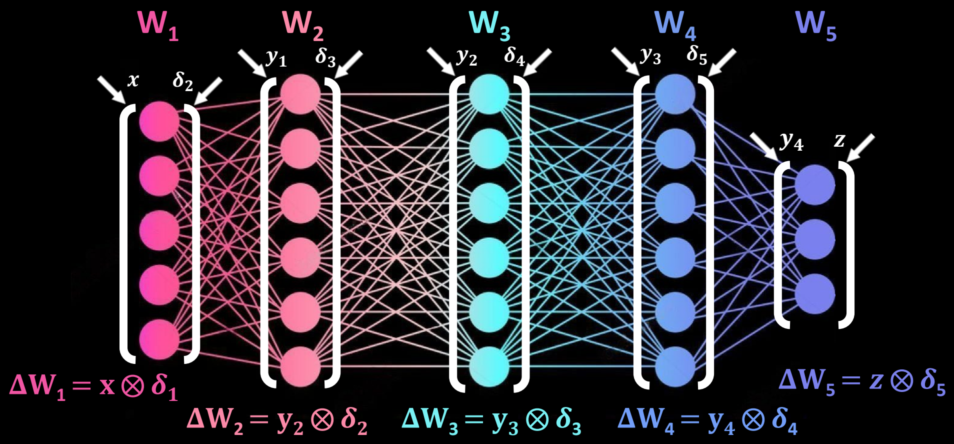
A. Reuther, et al., Survey of Machine Learning Accelerators, *HPEC,* (2020).

Deep Learning Operations

Testing Inputs (& Backpropagating Errors)



Updating Weights



Matrix Inner Products

Matrix Outer Products

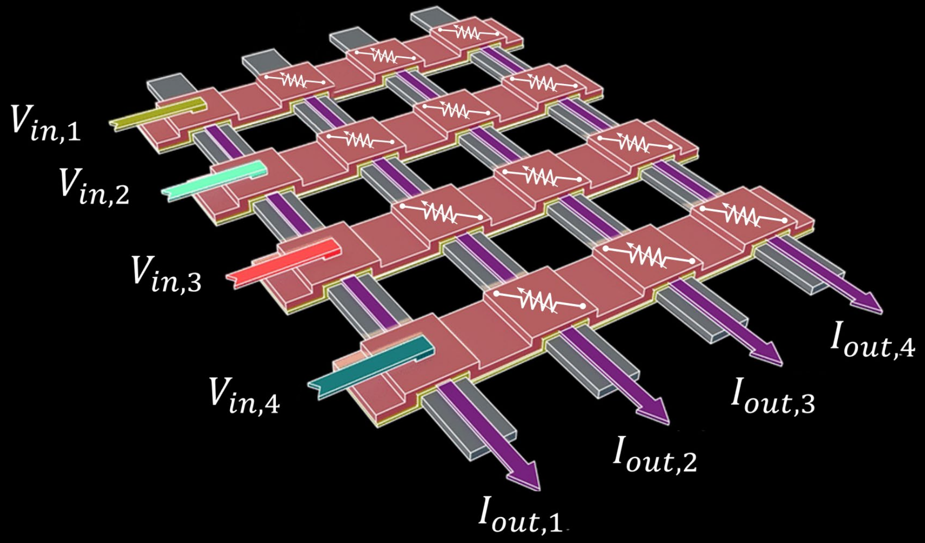
Need to carry matrices between memory and processor

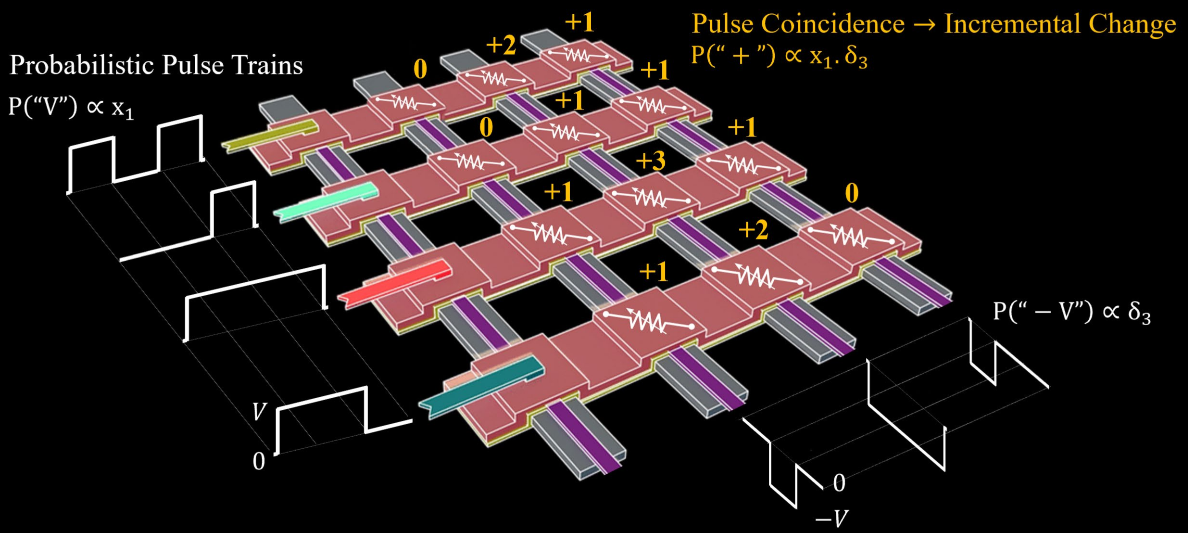
$\mathcal{O}(N^3)$ computational complexity

Analog Intrinsic Hardware for Deep Learning Operations

Testing Inputs (& Backpropagating Errors)

Updating Weights



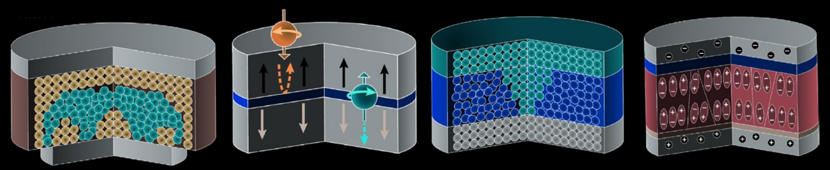


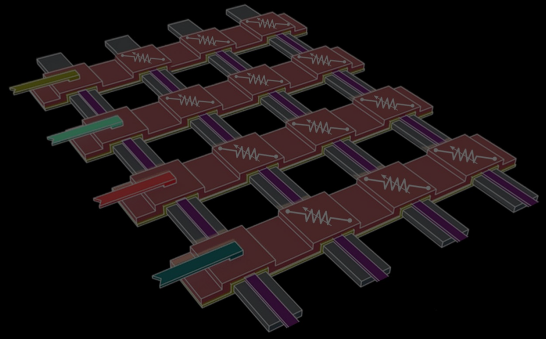
Matrix Inner Products

Matrix Outer Products

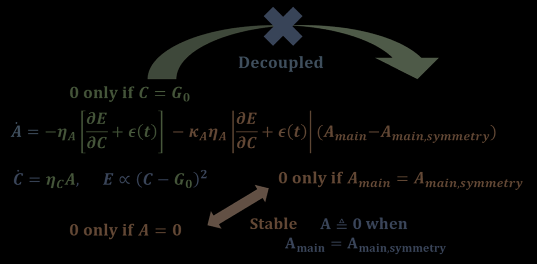
Local (in-memory) processing

$\mathcal{O}$(N) computational complexity (Fully-parallel operation)



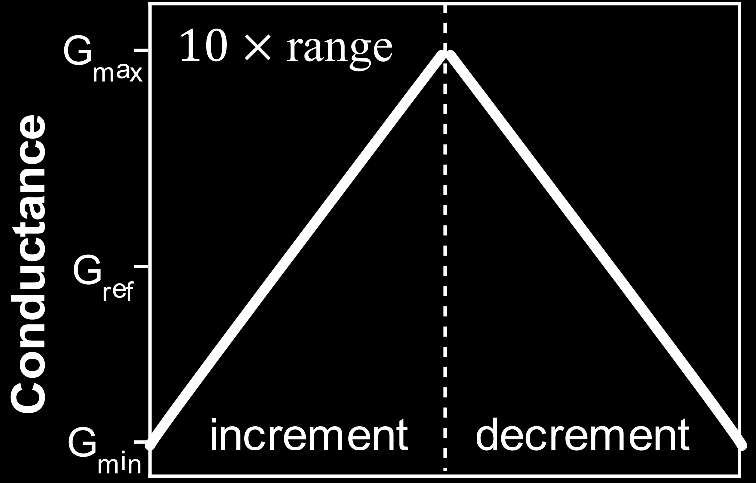


I. $y = Cx$
II. $z = C^T \delta$
III. $A \leftarrow A + \eta_A (x \times \delta)$
IV. $v = Au$
V. $C \leftarrow C + \eta_C (u \times v)$



Devices **Algorithms**

Ideal Nonvolatile Programmable Device

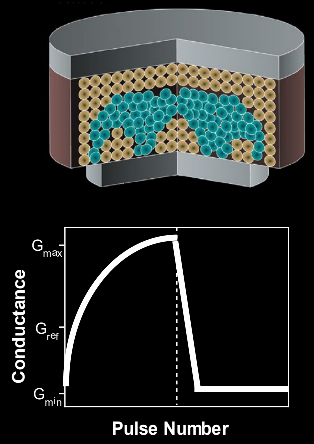


Pulse Number

↑↑↑ ••• ↑↑↓↓↓ ••• ↓↓

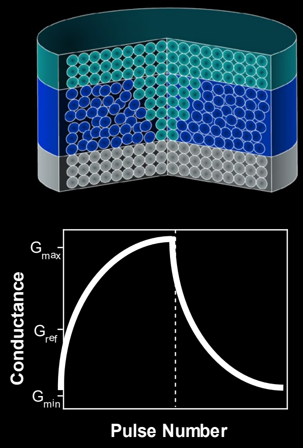
~1000 levels

Phase Change Memory



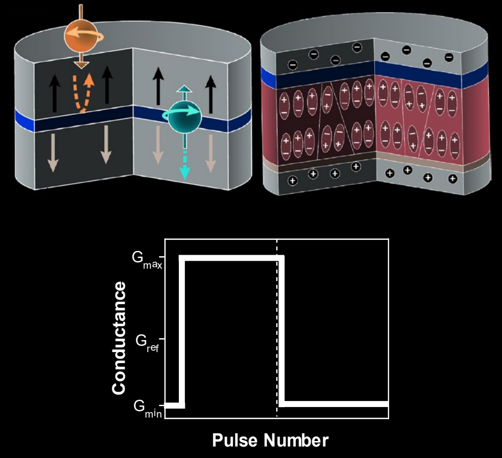
Abrupt Modulation

Filamentary RRAM



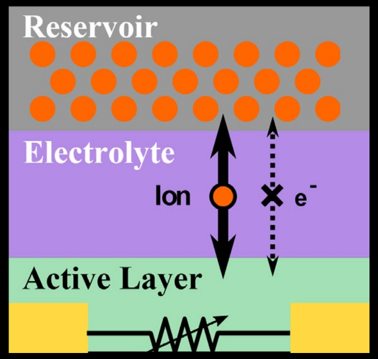
Asymmetric & Stochastic

Magnetic/Ferroelectric RAM



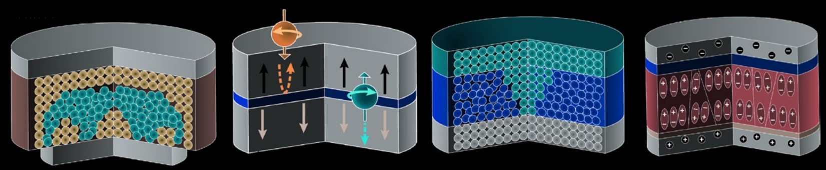
Uncontrollable & Binary

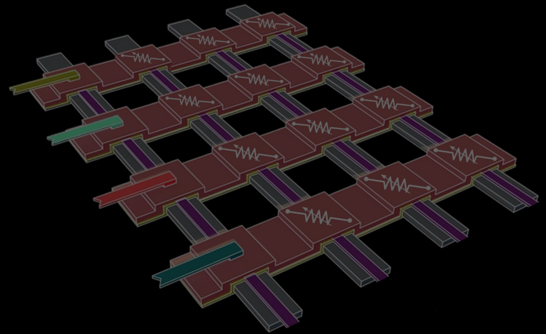
Non-filamentary Ionics



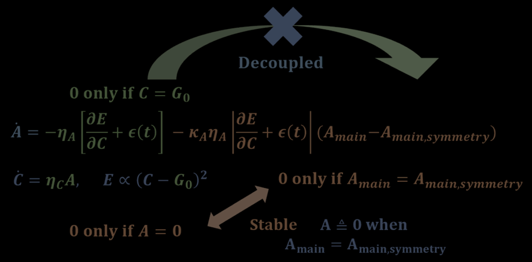
Use ions as dynamic dopants

T. Gokmen, Y. Vlasov, Acceleration of Deep Neural Network Training with Resistive Cross-Point Devices : Design Considerations, *Frontiers in Neuroscience,* (2016).
S. Agarval, et al., Resistive Memory Device Requirements for a Neural Algorithm Accelerator, *IJCNN,* (2016).

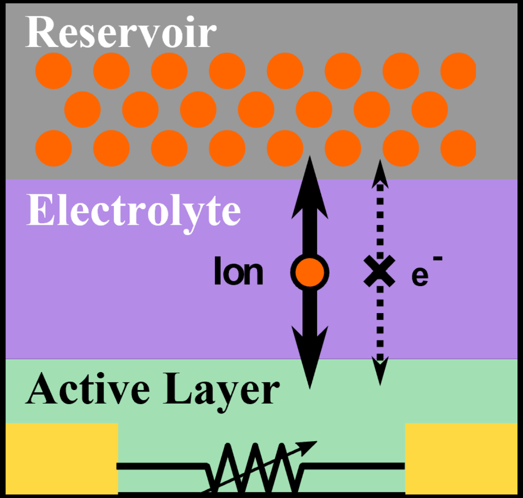


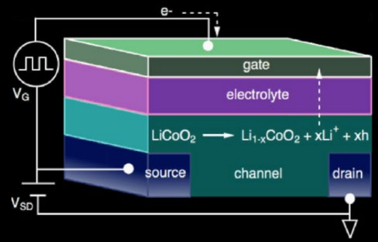


I. $y = Cx$
II. $z = C^T \delta$
III. $A \leftarrow A + \eta_A (x \times \delta)$
IV. $v = Au$
V. $C \leftarrow C + \eta_C (u \times v)$

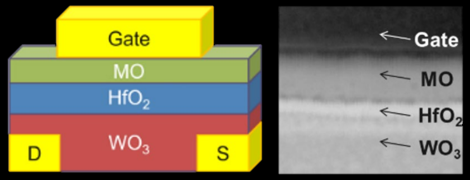


Li$^+$ Based | O^{2-} Based | H$^+$ Based

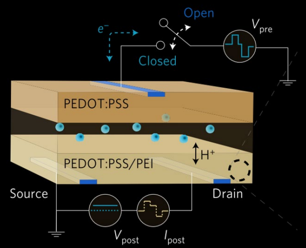




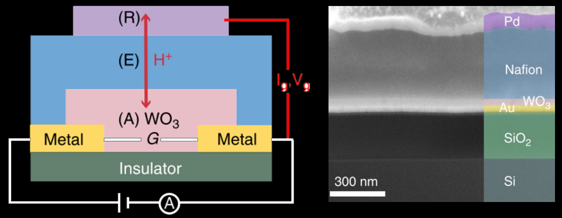
CMOS-incompatible



Slow Energy-inefficient



Liquid Electrolyte



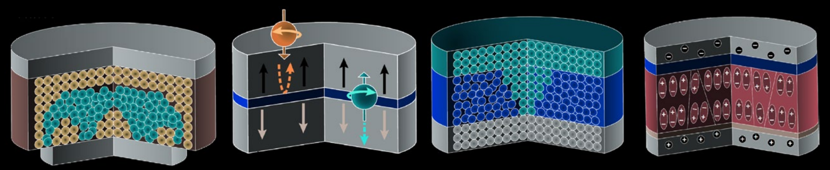
Polymer Electrolyte

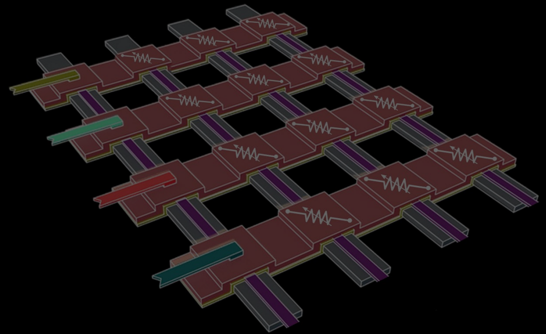
<u>Key Innovation:</u>

Phosphosilicate Glass (PSG, P-SiO$_2$)

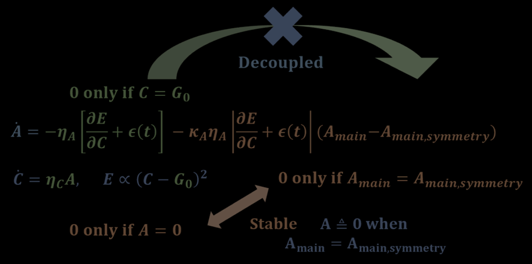
Si-compatible solid-state proton electrolyte

E. Fuller, et al, Li-Ion Synaptic Transistor for Low Power Analog Computing, *Advanced Materials,* (2017).
S. Kim, T. Todorov, <u>M. Onen</u>, et al., Metal-oxide based, CMOS-compatible ECRAM for Deep Learning Accelerator, *IEDM,* (2019).
Y. Van de Burgt, et al., A non-volatile Organic Electrochemical Device as a Low-voltage Artificial Synapse for Neuromorphic Computing, *Nature Materials,* (2017).
X. Yao, K. Klyukin, W. Lu, <u>M. Onen</u>, et al., Protonic Solid-State Electrochemical Synapse for Physical Neural Networks, *Nature Communications,* (2020).
<u>M. Onen</u>, et al., CMOS-compatible Protonic Programmable Resistor based on Phosphosilicate Glass Electrolyte for Analog Deep Learning, *Nano Letters,* (2021).
M. Roling, et al., Nonlinear ion transport in liquid and solid electrolytes, *EPJ,* (2017).



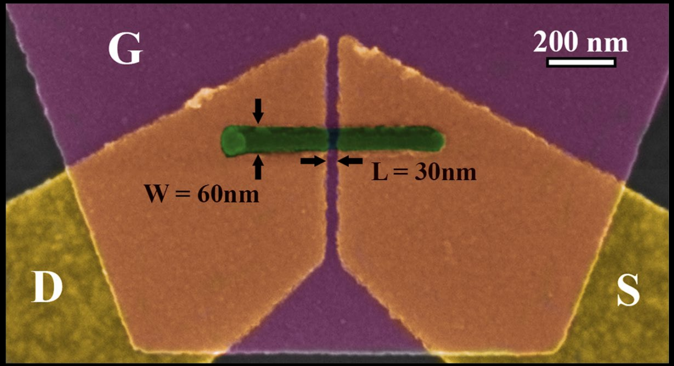


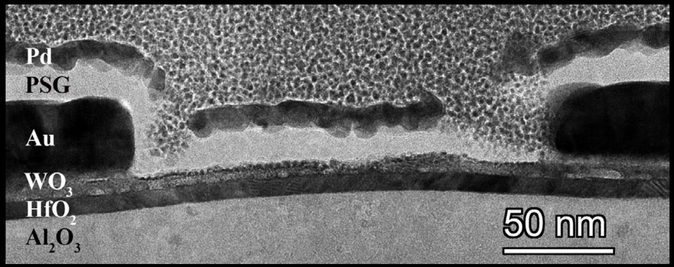
I. $y = Cx$
II. $z = C^T\delta$
III. $A \leftarrow A + \eta_A(x \times \delta)$
IV. $v = Au$
V. $C \leftarrow C + \eta_C(u \times v)$

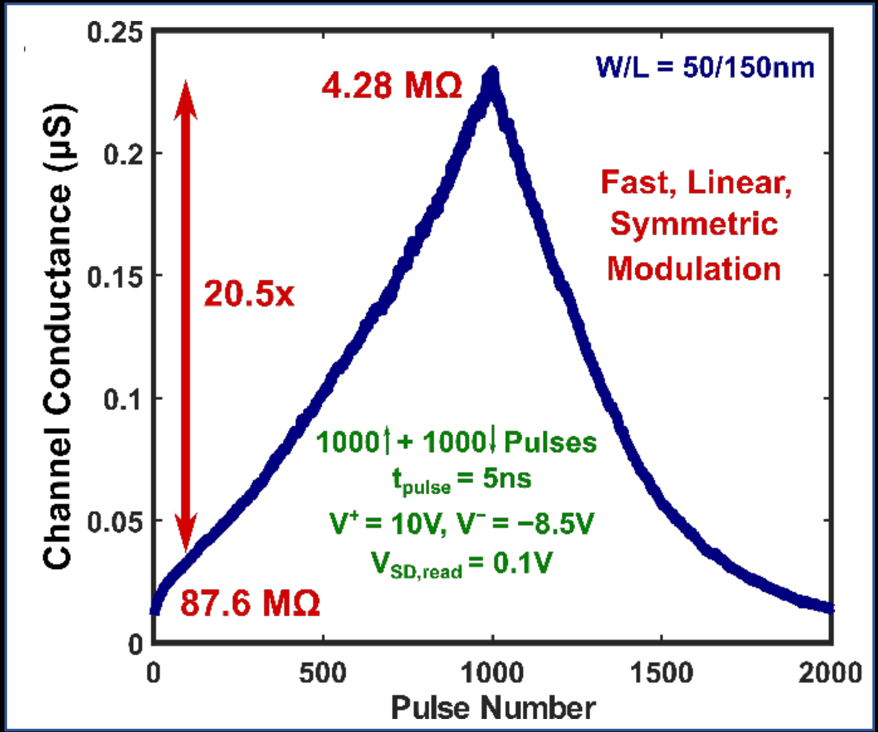


Devices

Algorithms



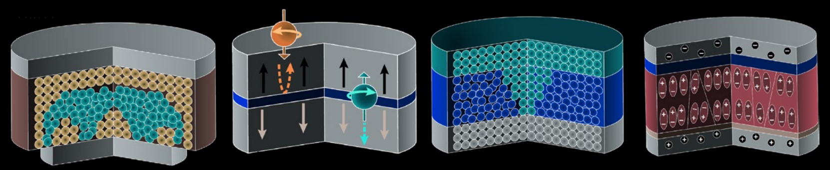


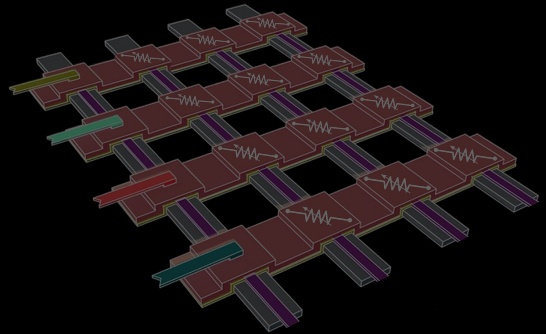


State-of-the-art Combined Properties

- ❖ Si-integration compatible
- ❖ Ultrafast ion transport (5 ns)
- ❖ High energy efficiency (2.5 fJ)
- ❖ Near-ideal modulation
- ❖ 1000 states across 20× range

M. Onen, et al., Nanosecond Protonic Programmable Resistors for Analog Deep Learning, *Science,* (2022).
M. Onen, et al., CMOS-compatible Protonic Programmable Resistor based on Phosphosilicate Glass Electrolyte for Analog Deep Learning, *Nano Letters*, (2021).
M. Onen, et al., Dynamics of PSG-Based Nanosecond Programmable Nanoprotonic Resistors for Analog Deep Learning, *IEDM*, (2022).





I. $y = Cx$
II. $z = C^T\delta$
III. $A \leftarrow A + \eta_A(x \times \delta)$
IV. $v = Au$
V. $C \leftarrow C + \eta_C(u \times v)$

Decoupled

0 only if $C = G_0$

$\dot{A} = -\eta_A\left[\frac{\partial E}{\partial C} + \epsilon(t)\right] - \kappa_A\eta_A\left|\frac{\partial E}{\partial C} + \epsilon(t)\right|(A_{main} - A_{main,symmetry})$

$\dot{C} = \eta_C A, \quad E \propto (C - G_0)^2$

0 only if $A_{main} = A_{main,symmetry}$

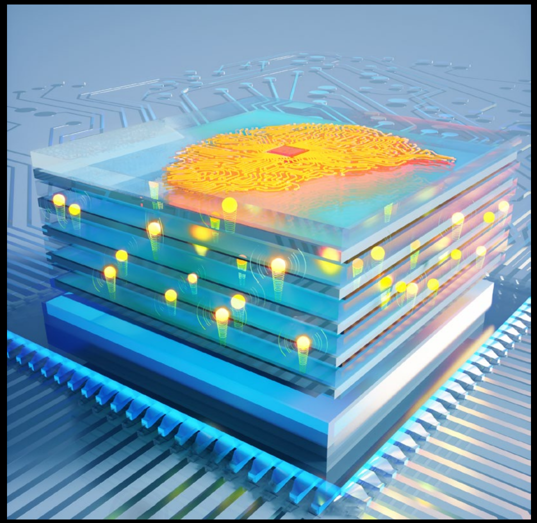
0 only if $A = 0$

Stable $\quad A \triangleq 0$ when $A_{main} = A_{main,symmetry}$

Devices — **Algorithms**

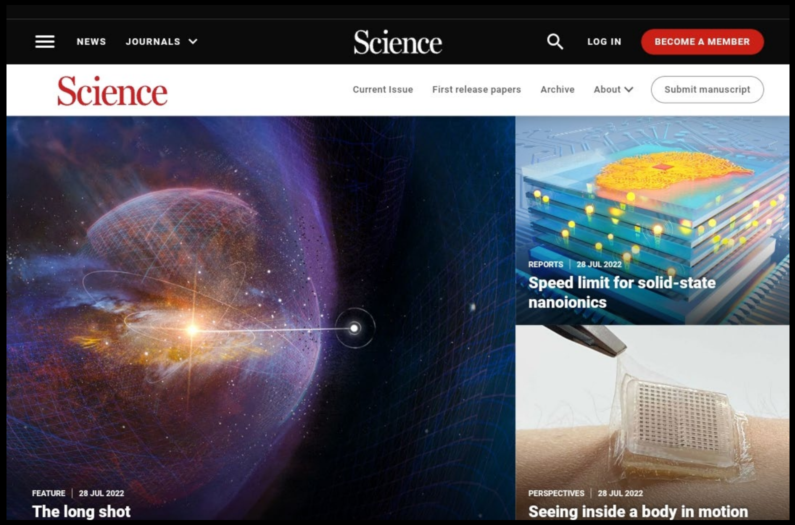
Science

July 29, 2022



M. Onen, et. al., Nanosecond Protonic Programmable Resistors for Analog Deep Learning, *Science,* (2022).

Highlighted on homepage

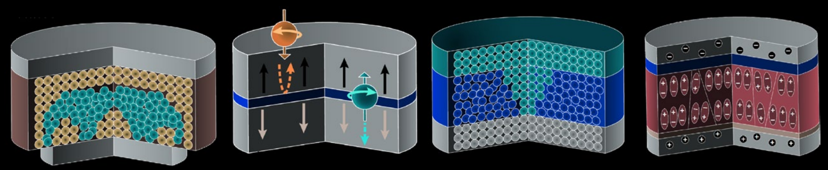


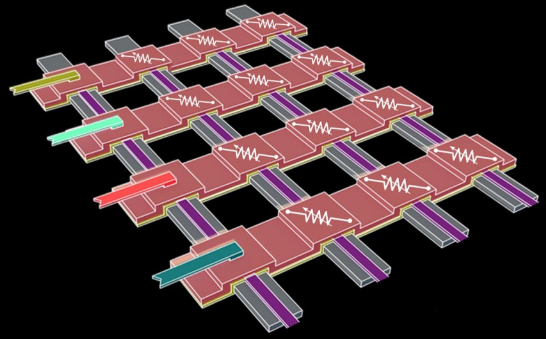
Press Coverage

electronicsforu.com, TECH TIMES, MIT News, Techxplore, BUG, Big News Network, Connect, Futurism, NOTIZIE IN, Enerzine, FUTUROPROSSIMO, tecnológica, GigaZINE, inSmart, cnBeta, newKerala.com, IMPARCIAL, Trust My Science, OSEL, TECHregister, elektroniknet.de, The Register, ANI News, EurekAlert!, labroots, ScienceDaily, The Brussels Times, msn, WebOrvos, fudzilla, Actu IA, News Break, ХАЙТЕК, LENTA.RU, UK Today News, SingularityHub, Smart Energy International, SecNews, Interesting Engineering, NEXT, EXBULLETIN, Tech News Space, TECHWORM, CIKAVOSTI, ThePrint, COSMOS, The Science Times, Mirage, DW, vosvete, sciencenewsnet.in, JdC, NewsDeal, LATESTLY, awsforwp, terra, IEEE Spectrum, New Atlas, WooDZog, CHIP, SwiftTelecast, inSmart, BarlettaWeb, Morning News, El Telégrafo, NewScientist, TN, The Byte, TechNewsBoy, haber7com, Thinking Port, ab-news.ru, Techxplore, Descoperă.ro, TIMES NOW, TweakTown, hackster.io, TimesNowNews.com

The New Yorker, The New York Times, NEO.LIFE, nature, nanotoday

Interviews to be published





Devices

I. $y = Cx$

II. $z = C^T \delta$

III. $A \leftarrow A + \eta_A (x \times \delta)$

IV. $v = Au$

V. $C \leftarrow C + \eta_C (u \times v)$

Decoupled

0 only if $C = G_0$

$$\dot{A} = -\eta_A \left[\frac{\partial E}{\partial C} + \epsilon(t)\right] - \kappa_A \eta_A \left|\frac{\partial E}{\partial C} + \epsilon(t)\right| (A_{main} - A_{main,symmetry})$$

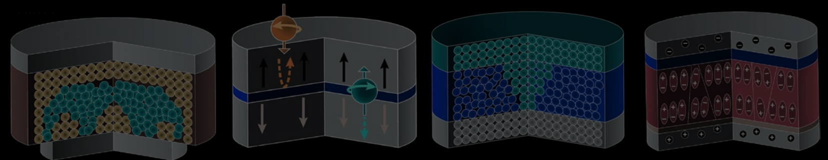
$$\dot{C} = \eta_C A, \quad E \propto (C - G_0)^2$$

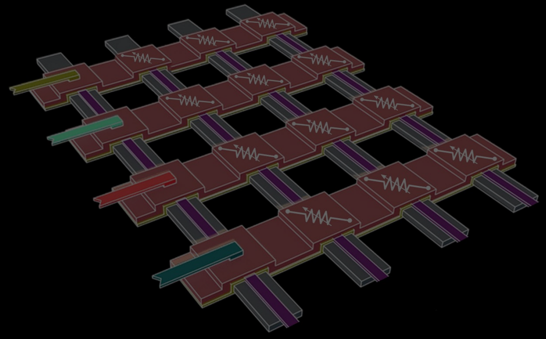
0 only if $A_{main} = A_{main,symmetry}$

0 only if $A = 0$

Stable $A \triangleq 0$ when $A_{main} = A_{main,symmetry}$

Hardware-aware training algorithms
Nonideality tolerance without compromising parallelism





I. $y = Cx$
II. $z = C^T\delta$
III. $A \leftarrow A + \eta_A(x \times \delta)$
IV. $v = Au$
V. $C \leftarrow C + \eta_C(u \times v)$

Decoupled

0 only if $C = G_0$

$$\dot{A} = -\eta_A\left[\frac{\partial E}{\partial C} + \epsilon(t)\right] - \kappa_A\eta_A\left|\frac{\partial E}{\partial C} + \epsilon(t)\right|(A_{main} - A_{main,symmetry})$$

$\dot{C} = \eta_C A, \quad E \propto (C - G_0)^2$

0 only if $A_{main} = A_{main,symmetry}$

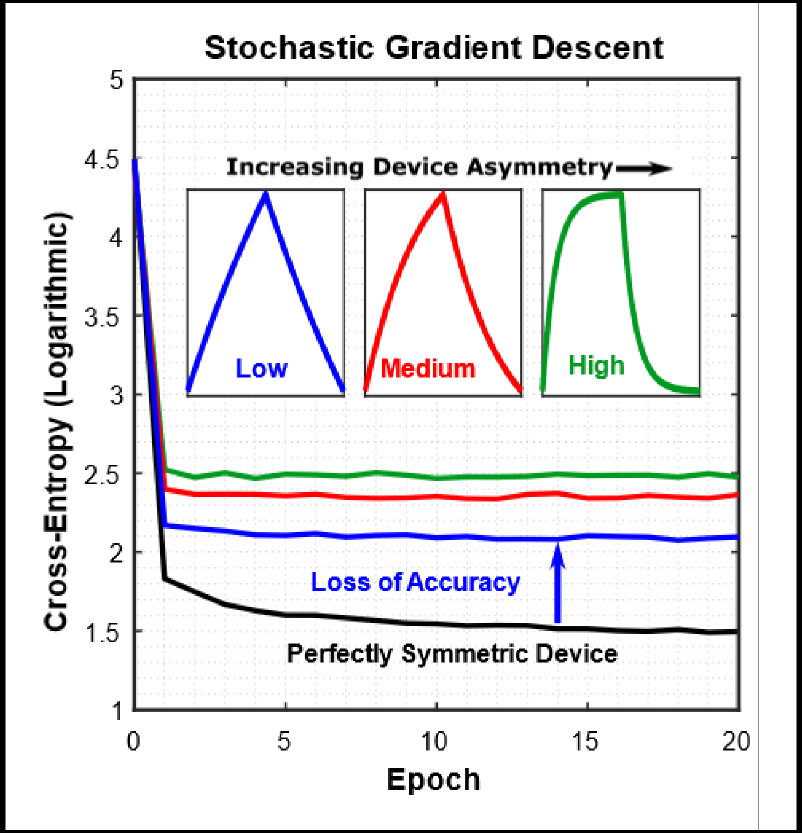
0 only if $A = 0$

Stable

$A \triangleq 0$ when $A_{main} = A_{main,symmetry}$

Devices

Algorithms



The Infamous Device Asymmetry

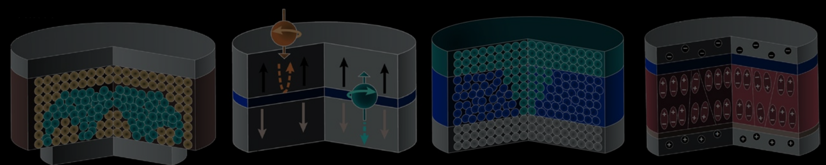
Proposed Solutions:

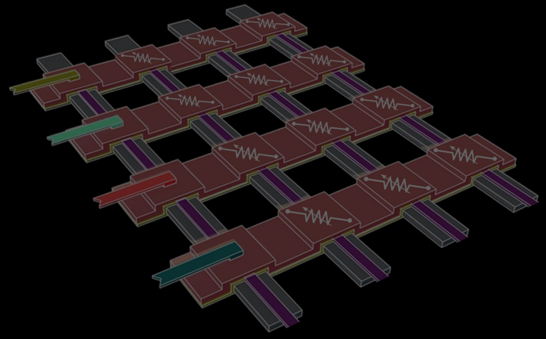
- ❖ Read-write-verify
- ❖ Weight carry

Sacrifices parallelism

- ❖ Individual pulse-control
- ❖ Compliance circuitry
- ❖ Multiple device-copy

Sacrifices area & energy-efficiency





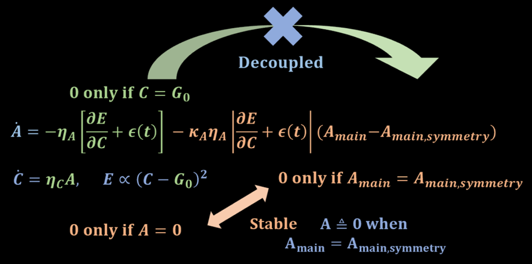
I. $y = Cx$

II. $z = C^T\delta$

III. $A \leftarrow A + \eta_A(x \times \delta)$

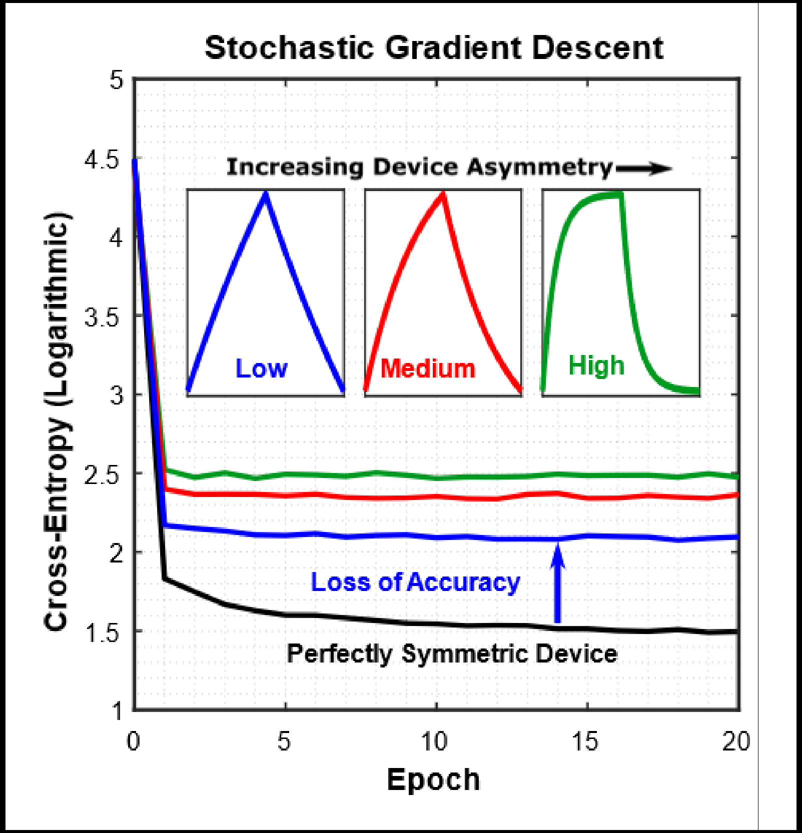
IV. $v = Au$

V. $C \leftarrow C + \eta_C(u \times v)$



Devices

Algorithms

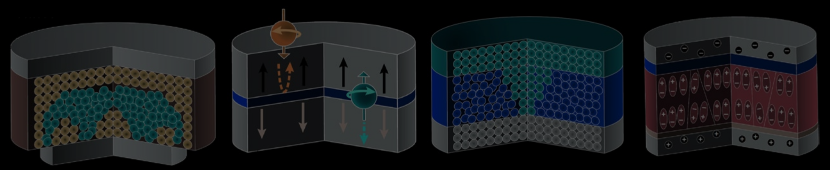


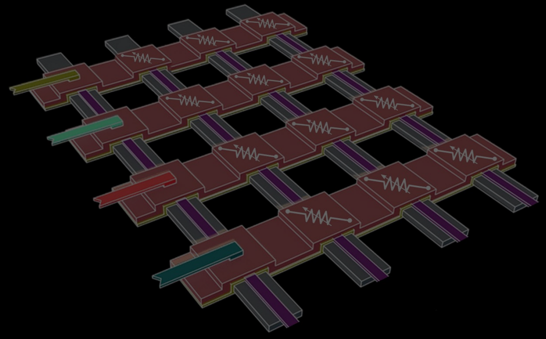
A More Fundamental Problem…

SGD Updates: Gradient of error (potential energy)
Points towards minima

Asymmetry: Error with each change (kinetic energy)
Points to the middle (competing minima!)

Incorporate hardware energy to the error function





I. $y = Cx$
II. $z = C^T\delta$
III. $A \leftarrow A + \eta_A(x \times \delta)$
IV. $v = Au$
V. $C \leftarrow C + \eta_C(u \times v)$

Decoupled

0 only if $C = G_0$

$$\dot{A} = -\eta_A\left[\frac{\partial E}{\partial C} + \epsilon(t)\right] - \kappa_A\eta_A\left|\frac{\partial E}{\partial C} + \epsilon(t)\right|(A_{main} - A_{main,symmetry})$$

$$\dot{C} = \eta_C A, \quad E \propto (C - G_0)^2$$

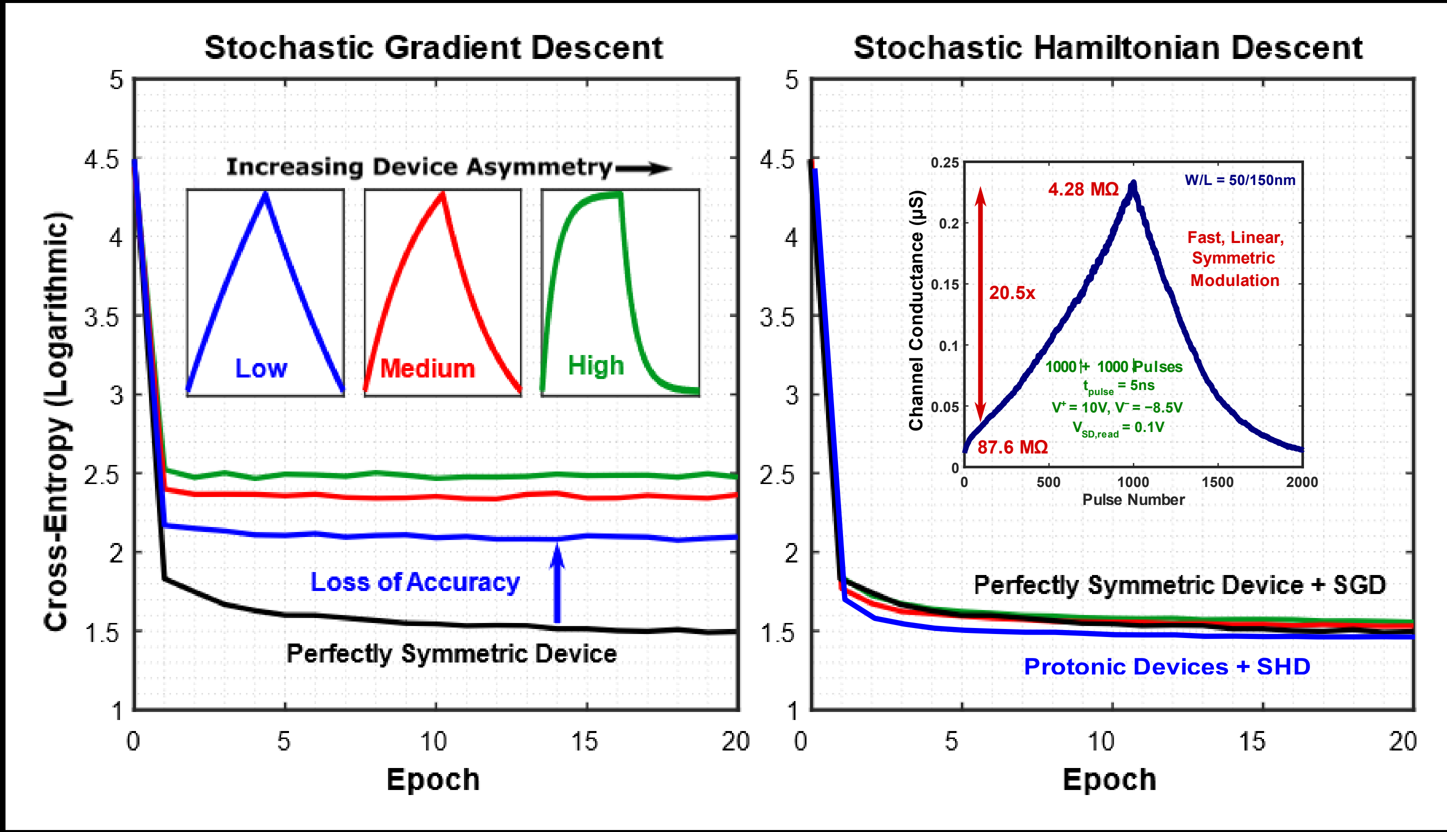
0 only if $A_{main} = A_{main,symmetry}$

0 only if $A = 0$

Stable $A \triangleq 0$ when $A_{main} = A_{main,symmetry}$

Devices

Algorithms



Advanced Training Algorithm: SHD

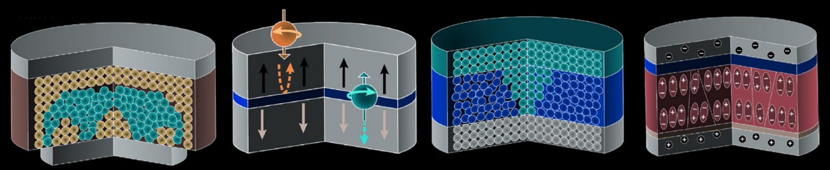
- ❖ Resolves infamous asymmetry penalty
- ❖ Preserves fully-parallel operation
- ❖ SHD + protonic devices > SGD + "ideal"

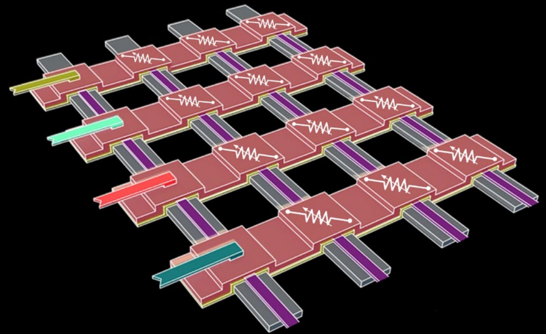
Intrinsic Processors: Beyond Deep Learning

- ❖ Matrix sketching & streaming
- ❖ Randomized numerical algorithms
- ❖ Generalized linear algebra

M. Onen, et. al., "Accelerated Quasi-Newton Methods on Analog Crossbar Hardware", Pending, US/2022/0083623.
M. Onen, et. al., "Matrix Sketching Using Analog Crossbar Architectures", *Pending*, US/2021/0357540, WO/2021/229320

M. Onen, et.al., "Neural Network Training with Asymmetric Crosspoint Elements", Frontiers in Artificial Intelligence, (2022).





I. $y = Cx$

II. $z = C^T \delta$

III. $A \leftarrow A + \eta_A (x \times \delta)$

IV. $v = Au$

V. $C \leftarrow C + \eta_C (u \times v)$

Decoupled

0 only if $C = G_0$

$$\dot{A} = -\eta_A \left[\frac{\partial E}{\partial C} + \epsilon(t)\right] - \kappa_A \eta_A \left|\frac{\partial E}{\partial C} + \epsilon(t)\right| (A_{main} - A_{main,symmetry})$$

$\dot{C} = \eta_C A, \quad E \propto (C - G_0)^2$

0 only if $A_{main} = A_{main,symmetry}$

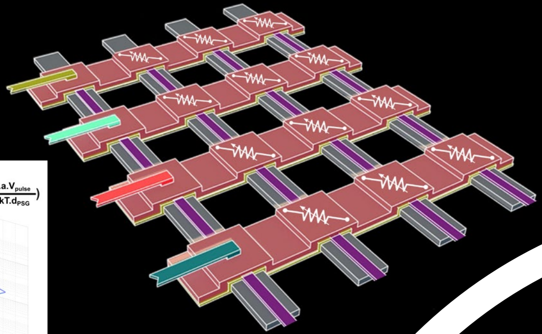
0 only if $A = 0$

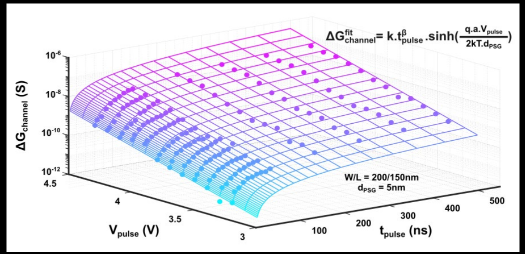
Stable $\quad A \triangleq 0$ when $A_{main} = A_{main,symmetry}$

Devices — Architectures — Algorithms

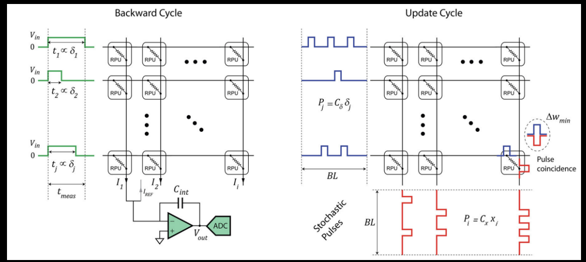
How to innovate faster?

Co-optimization.



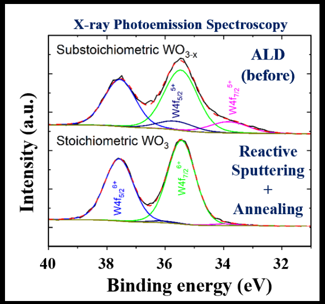


Efficient Architectures

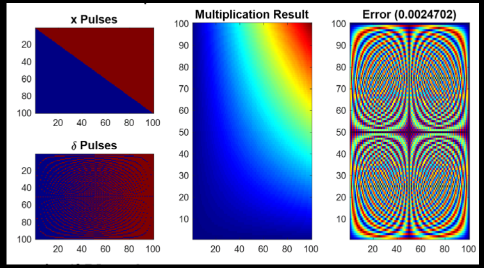


Circuit Optimization
System Pipelining
Analog-Digital Interfacing

Material Science
Device Physics
Nanofabrication

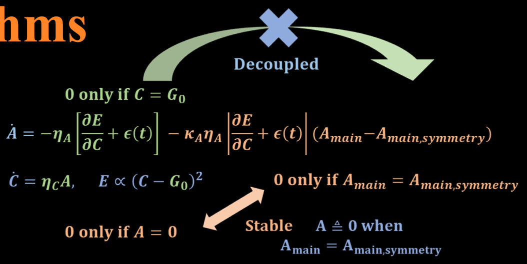


Co-optimization.

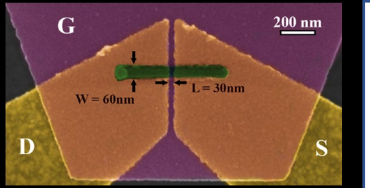


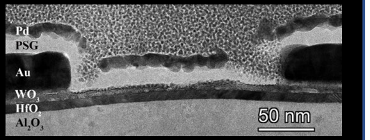
Analog-Specific Algorithms

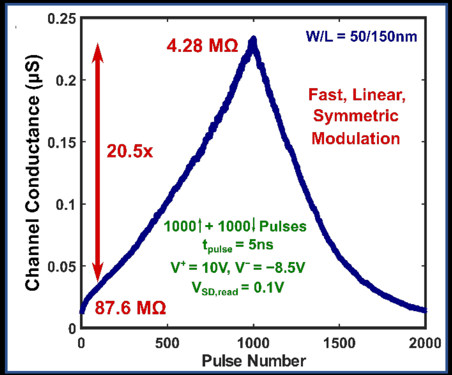
I. $y = Cx$
II. $z = C^T\delta$
III. $A \leftarrow A + \eta_A(x \times \delta)$
IV. $v = Au$
V. $C \leftarrow C + \eta_C(u \times v)$

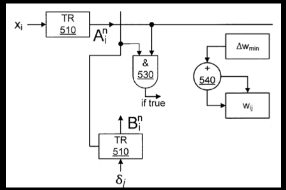


Optimal Devices









Device Modeling
Applied Mathematics
Fault-tolerant Computing

Devices and Algorithms for Analog Deep Learning

Presenter: Murat Onen
monen@mit.edu

Presented to: MIT AI Hardware Program
Cambridge, MA
10/14/2022

Research Team: Jesús A. del Alamo
(devices) Bilge Yıldız
Ju Li
Frances M. Ross
Nicolas Émond
Baoming Wang
Difei Zhang

Research Sponsor: MIT-IBM Watson AI Lab

Acknowledgements: MIT, IBM, intel, DARPA, Raytheon BBN

Collaborators:

Mantao Huang	Tayfun Gokmen	Teodor K. Todorov	Vasileios Kalantzis
Miranda Schwacke	Malte Rasch	Paul Solomon	Haim Avron
Alexandre C. Foucher	Lior Horesh	Guy Cohen	Seyoung Kim
Seungchan Ryu	Wilfried Haensch	John Rozen	Tomasz Nowicki