Perspectives and Opportunities in AI Hardware

Jeff Burns
Director, AI Compute and IBM Research AI Hardware Center

November 5, 2021

© 2021 IBM Corporation

The Future of Computing

Bits

Mathematics + Information

Today's Computers and Supercomputers

Neurons

Biology + Information

Today's AI Systems

Qubits

Physics + Information

Today's Quantum Systems

© 2021 IBM Corporation

The evolution of AI

We are here

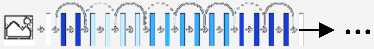
Narrow AI	**Broad AI**	**General AI**
Deep learning	Learning + reasoning	True neuro-AI
Single-task, single-domain, with superhuman accuracy	Multi-task, multi-domain, multi-modal	Cross-domain learning and reasoning
Requires large amounts of labeled data	Learns with much less data	Broad autonomy

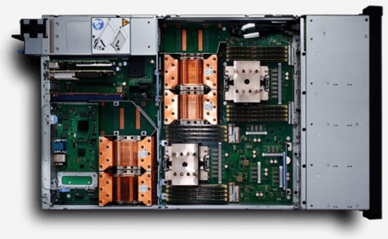
Even “narrow AI” relies on computation horsepower

Training Image recognition model

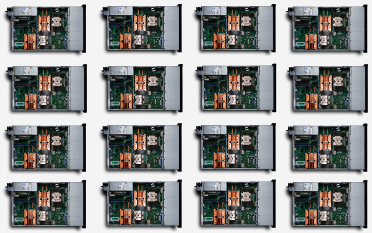
Dataset:
ImageNet-22K

Network:
ResNet-101





4 GPUs
16 days
~385 kWh



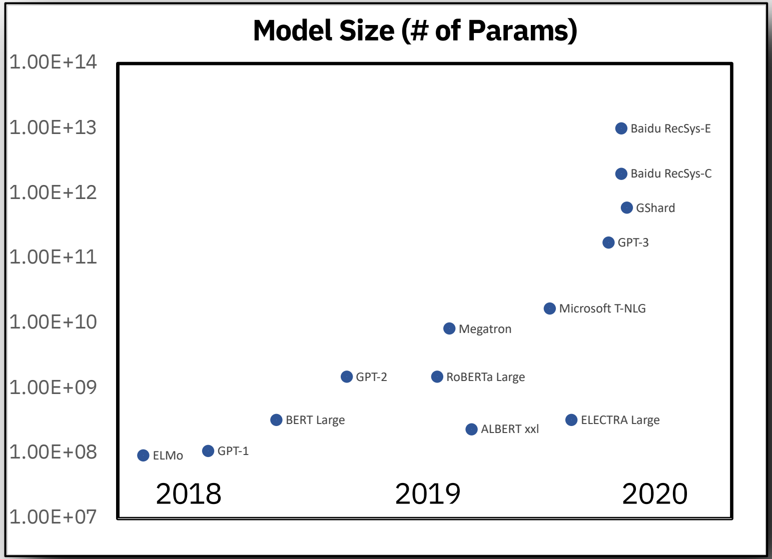
256 GPUs
7 hours
~450 kWh

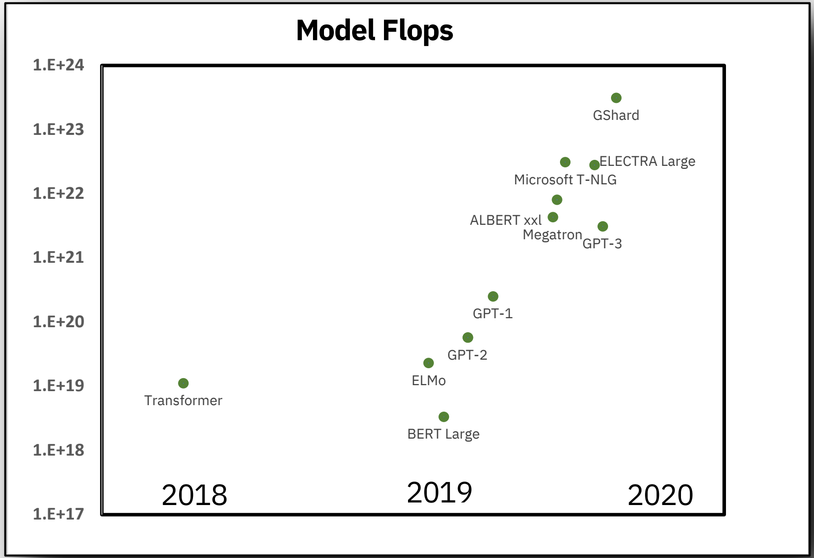
1 model training run is ~2 weeks of home energy consumption

https://arxiv.org/abs/1708.02188

IBM Research AI Hardware Center / © 2021 IBM Corporation

Explosive Growth in AI Compute Needs

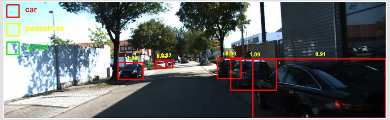




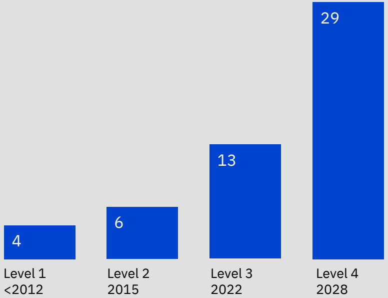
- Artificial Intelligence is being applied to an increasing number of domains (vision, speech, NLP ...)
- Explosive growth in model sizes and flops over the past 3-5 years (especially in NLP, recommender, graph models)
- AI accelerator performance needs to grow exponentially to keep up with model growth

“Broad AI” brings even more computational demands and greater functionality requirements at the edge

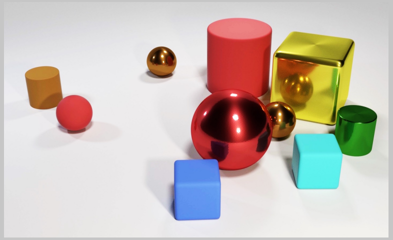
Multi-Modal Models



Number of sensors for different levels of autonomous driving (source: Deloitte)



Explainability with Neuro-Symbolic Reasoning

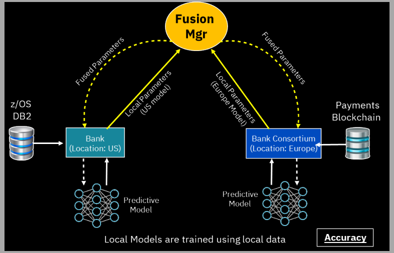


Question: *Are there an equal number of large things and metal spheres?*

Program: equal number (count(filter_size(Scene, Large)), count(filter_material (filter_shape(Scene, Sphere), Metal)))

Answer: *Yes*

Security and Privacy



Federated learning, data stays at the edge

IBM Research AI Hardware Center

"IBM invests $2 Billion in New York Research Hub for AI"

Bloomberg

"IBM Bets $2B Seeking 1000X AI Hardware Performance Boost"

insideHPC.

An ecosystem of enterprise and academic partners

February 7, 2019
Launch Date

$2B
IBM Investment To Create Artificial Intelligence Hardware Center

$300M
New York State investment

17 and growing
Members of the IBM Research AI Hardware Center

© 2021 IBM Corporation

IBM Research AI Hardware Center

Challenge and Opportunity
AI present an incredible opportunity to extend automation – but at dramatic computational cost

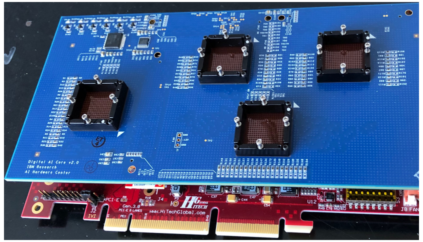
Objective
Innovate and lead in AI accelerators for training and inferencing

Technical Approach
Drive leadership using a full-stack strategy, generating AI accelerator demonstrators with an industry leading roadmap

Partnership
Engage partners to build a community and ecosystem to enable broad application of the Center's innovations

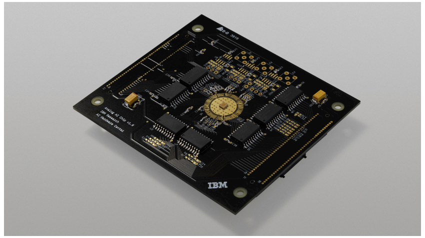
Cores and Architecture

New digital AI cores and architectures, based on fundamental algorithm and computational innovations



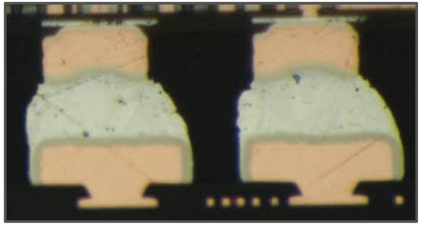
Analog Elements

Materials and architectural innovations to enable analog computation for AI inference and training



Heterogeneous Integration

Innovations in advanced laminate, silicon bridges, and 3D to scale connectivity and mitigate bandwidth bottlenecks



End User AI Testbed

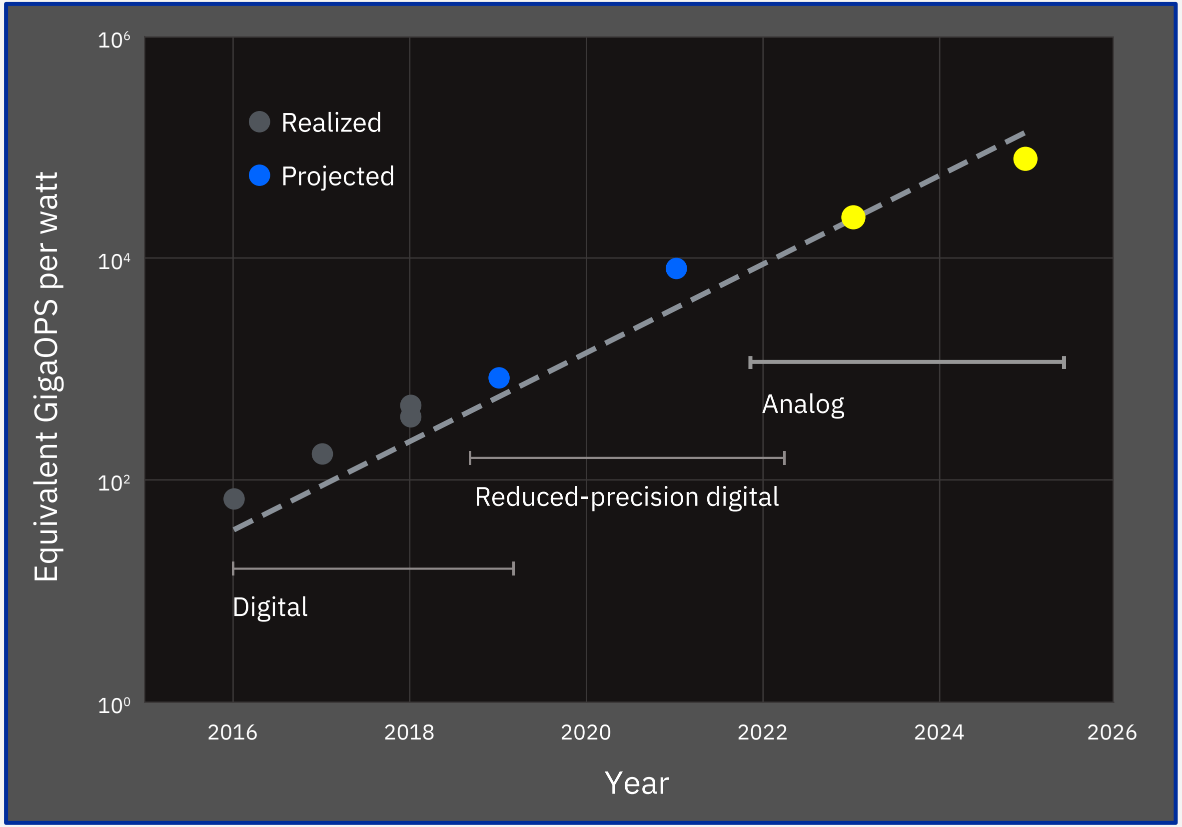
Leverage and develop advanced AI software to utilize new accelerators and capture emerging workload needs

What’s next in AI hardware

Extending performance by 2.5X / year through 2025

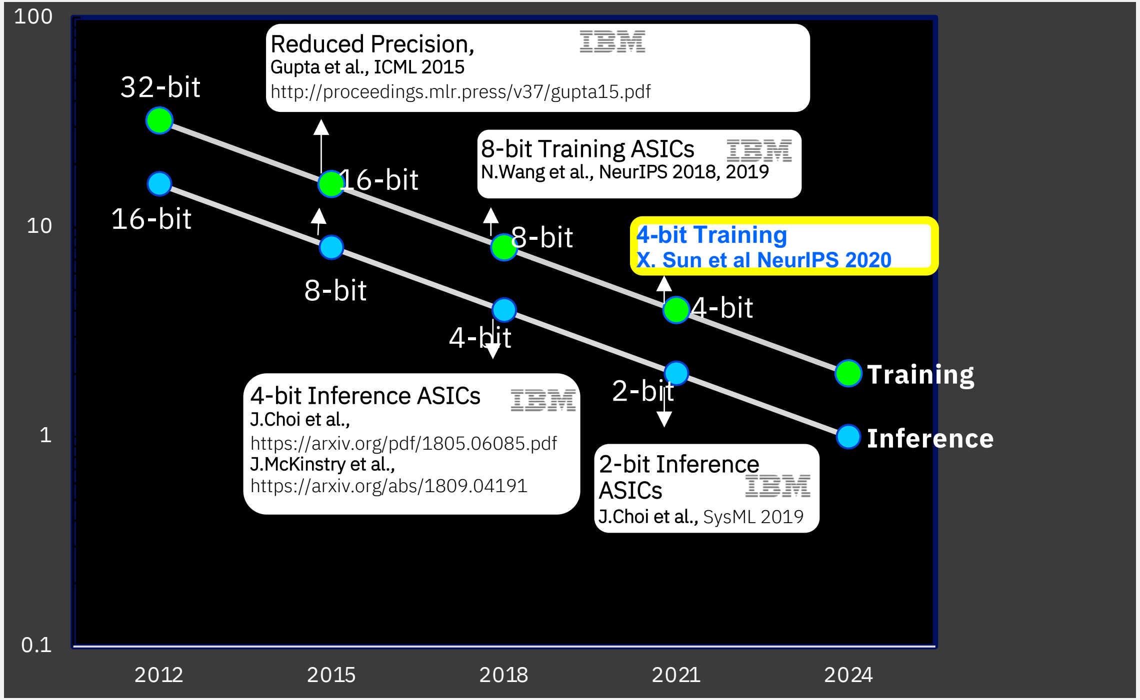
Approximate computing principles applied to **Digital AI Cores** with reduced precision, as well as

Analog AI Cores,
which could potentially offer another
100x in energy-efficiency



T. Gokmen and Y. Vlasov, *Frontiers in Neuroscience* **10**, pp. 333, 2016

Driving reduced precision *with iso accuracy*

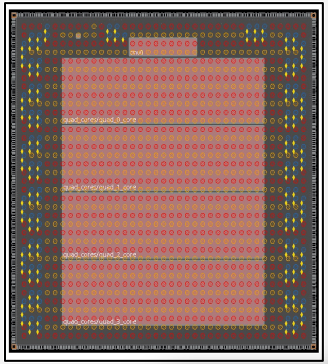


- Key advancements in reduced precision arithmetic for AI driven by IBM AI Research team.
- First demonstration of 16-bit precision for Deep Learning Training (ICML 2015).
- Demonstration of world's first 8-bit training (NeurIPS 2018, NeurIPS 2019), and world's first 4-bit training (NeurIPS 2020).
- Demonstration of highly accurate 2-bit and 4-bit Inference (SysML 2019)

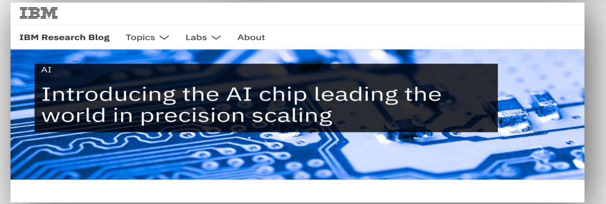
For reference - Industry standard for training:
- GPU default: 32 bit
- GPU accelerated: 16 bit (V100 & A100)
- TPU: 16 bit (Bfloat)

Digital AI core innovations

100X Improvement in 3 Years!

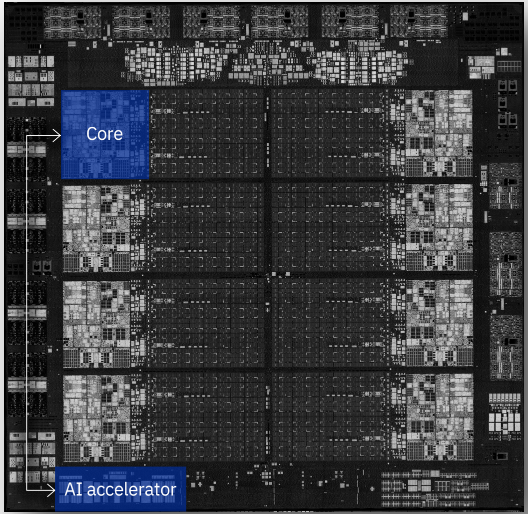


Precision	Power-efficiency
fp16 (T, I)	>2.5 TOPs/W
fp8 (T, I)	**> 5.5 TOPs/W**
int4 (I)	**> 20 TOPs/W**
int2 (I)	> 40 TOPs/W


IBM
IBM Research Blog Topics Labs About
AI
Introducing the AI chip leading the world in precision scaling


The Linley Group
MICROPROCESSOR *report*
Insightful Analysis of Processor Technology
IBM DEMONSTRATES NEW AI DATA TYPES
Research Chip Proves Value of FP8 Training, INT4 Inference
By Linley Gwennap (April 5, 2021)

Next generation Z processor is optimized to run enterprise workloads with **embedded real time AI insights**



AI Specifications

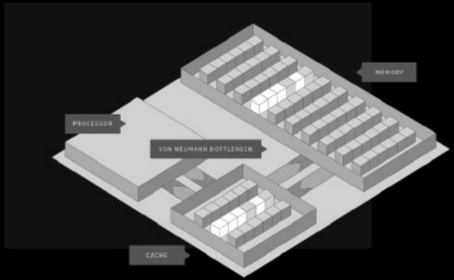
- **6 TFlops/chip**
- Up to 200 TFlops/syste m
- Focused on **low-latency** AI Inference


IBM Telum – A New Chapter In Vertically Integrated Chip Technology
Patrick Moorhead Senior Contributor
Cloud
I write about disruptive companies, technologies and usage models.

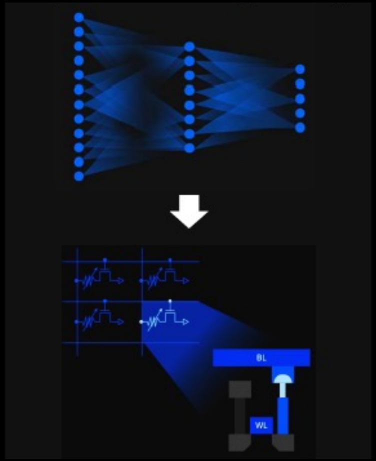
Analog NVM for in-memory compute

Eliminate the Von-Neumann bottleneck

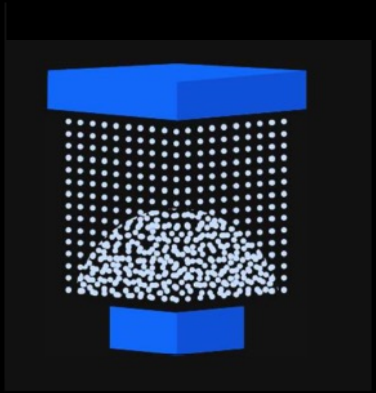
Perform computation directly in memory



Map DNNs to analog cross-point arrays



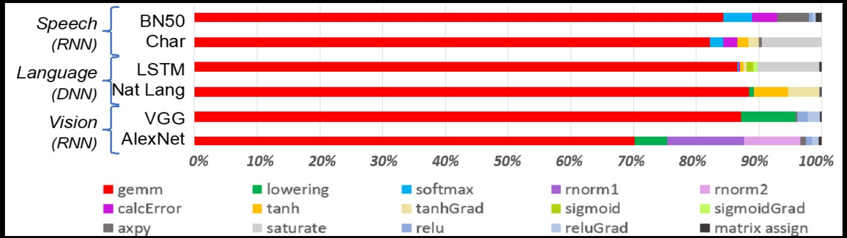
NVM materials in array crosspoints to store weights



© 2021 IBM Corporation

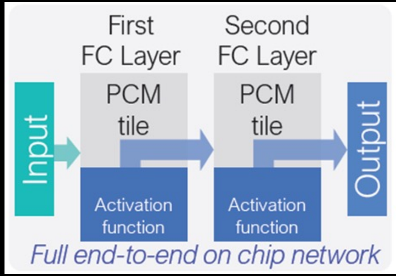
Key advantages of analog AI inference

- **Improved energy efficiency**
  - Significantly higher power efficiency for in-memory MAC compute (DL Inference dominated by MAC ops)



- **Zero standby power (leakage)**
  - Takes advantage of non-volatile memory technology
  - Low start-up time (no need to fetch the weights from memory)

- **Very low latency**
  - Takes advantage of pipelined ‘weight stationary’ architecture
  - Latency ≤ 1 msec for most models/workloads
  - Advantageous for low mini-batch ‘streaming’ workloads

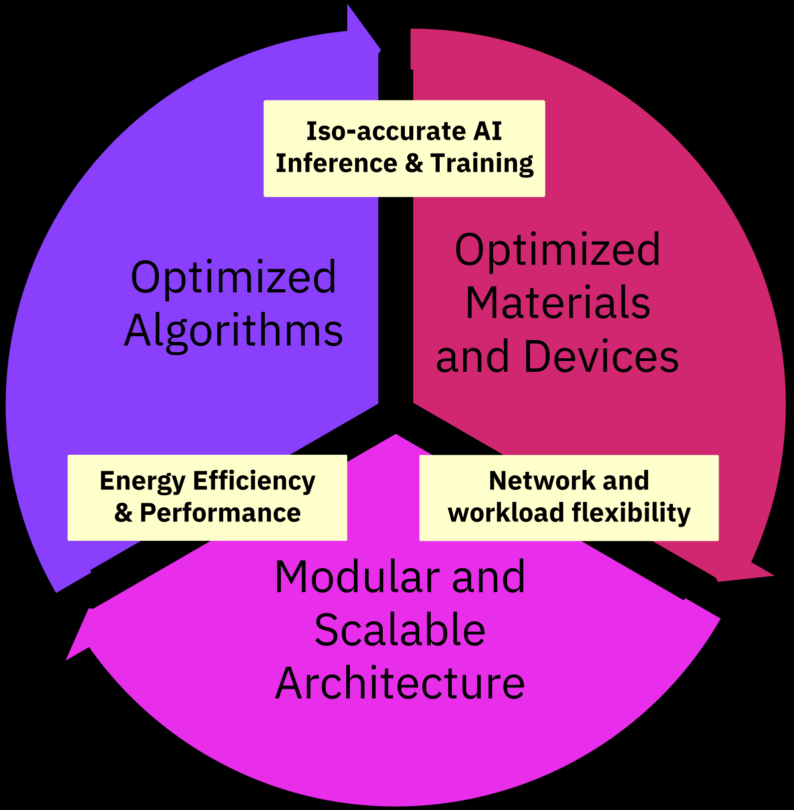


What should be attributes of an analog AI accelerator?

- ➢ Iso-accurate AI Inference and Training across multiple networks and workloads

- ➢ Flexible and modular architecture to scale to larger models

- ➢ Technology, algorithms & architecture for energy efficiency and performance



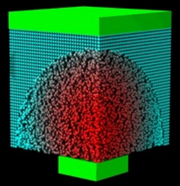
Materials/device requirements for AI inference

Forward inference

(Fixed weight)

Long-term retention
Excellent conductance stability
(Non-idealities: Drift, Noise, Stochasticity & Temp variations)
Modest endurance
Modest programming speed

Phase change memory (PCM)
e.g. $Ge_2Sb_2Te_5$



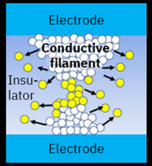
Training

(Frequent weight updating)

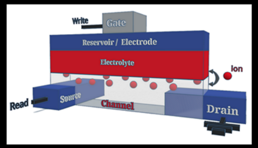
Modest retention
High endurance
Fast programming speed
Symmetric & gradual conductance change*

**Algorithmic innovation has mitigated need for symmetric update*

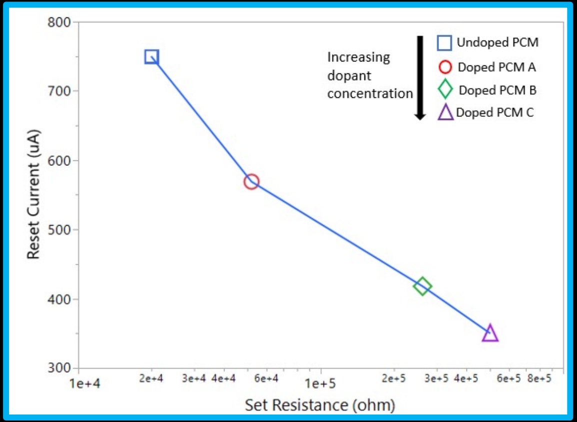
Resistive RAM (RRAM)
e.g. HfO_2

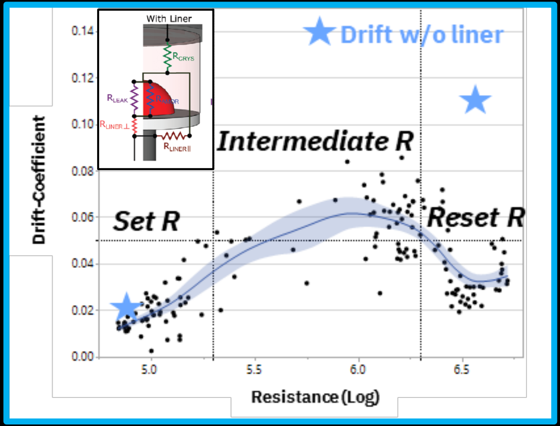


Electro-chemical RAM (ECRAM)
e.g. HfO_2 on WO_3 channel



PCM materials & device improvements



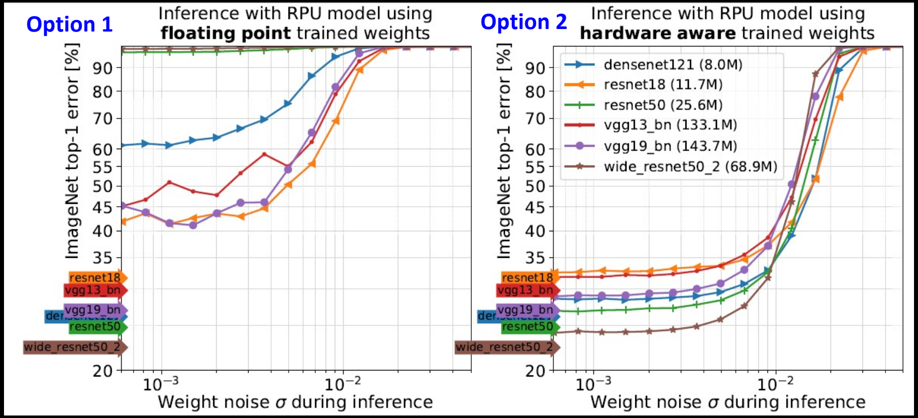


- Doped phase change materials for optimized device characteristics
  - Materials optimized to meet SET resistance and RESET current requirements

- Optimized projection liner for reduced drift-coefficient
  - Significant reduction in resistance-drift coefficient at RESET state
  - Also, reduced drift coefficient in intermediate resistance states

The path to ideal analog compute

Algorithmic Boosters: Hardware-aware (re)training for ‘Iso-accurate’ Analog AI Inference



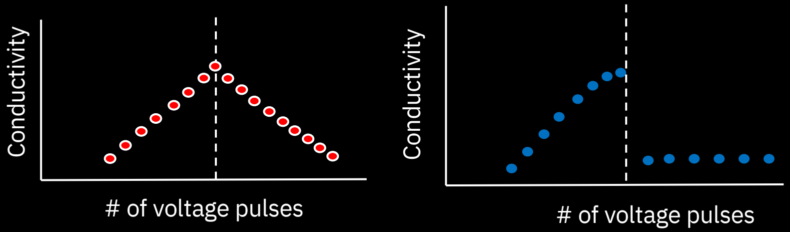
J.P. Han et al, SSDM, 2020

- Incorporate analog deficiencies & non-idealities (noise, circuit offsets, ADC/DAC resolutions etc) into the forward training pass
- Re-training in a hardware-aware (HWA) fashion increases robustness of inference to analog NVM and peripheral circuit non-idealities
- Near Iso-accurate inference performance achieved for a variety of DNNs (CNNs, LSTM, Transformer) & workloads (NLP, Speech, Image)

© 2021 IBM Corporation

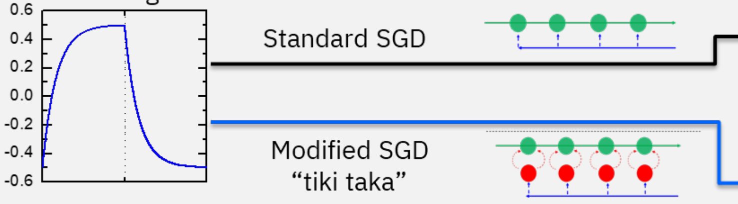
The path to ideal analog compute

Algorithmic Boosters: Algorithmic correction of asymmetry for Training

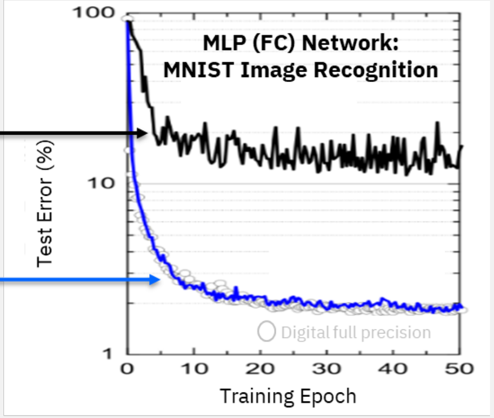


Material with asymmetric switching

Training with pulse train of **constant width & amplitude**

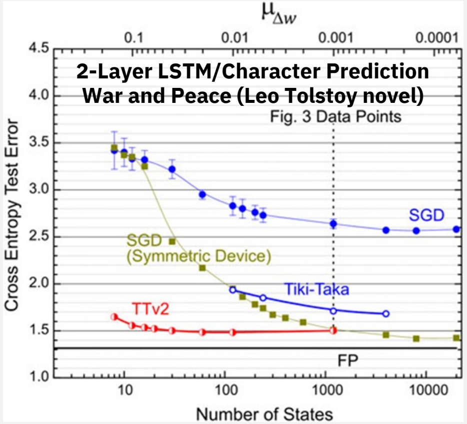


Algorithmic compensation of asymmetry achieves same accuracy as full-precision digital



T. Gokmen, and W.Haensch, Front. Neurosci., 26 February 2020 | https://doi.org/10.3389/fnins.2020.00103

Improved Training Algorithm - helps ease stringent device requirements for more complex models



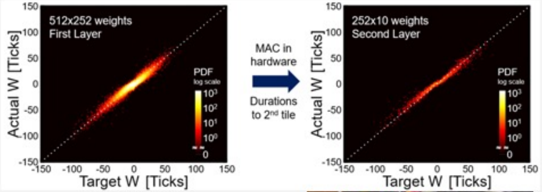
T. Gokmen et al., Front. Neurosci, 4:126 (2021)

- *Algorithmic innovation has mitigated need for symmetric updates*
- *Continued improvements in the training algorithm has helped ease stringent device requirements for number of states & read noise*

© 2021 IBM Corporation

Inference: achievements to date

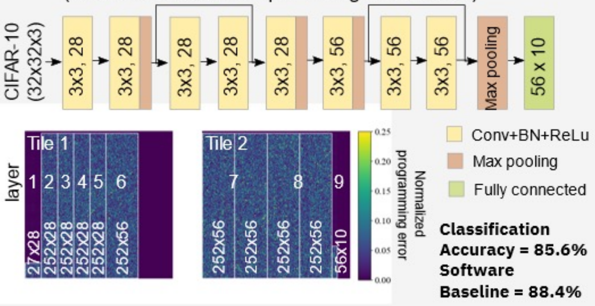
1 & 2 Layer MLP (FC)
MNIST
Image Recognition



Inference Accuracy	Hardware Measurements	Software Baseline
Design 1	97.12	97.73
Design 2	98.3	98.6

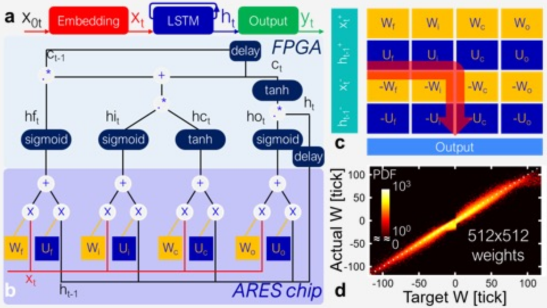
RESNET-9 (CNN)
CIFAR-10
Image Recognition

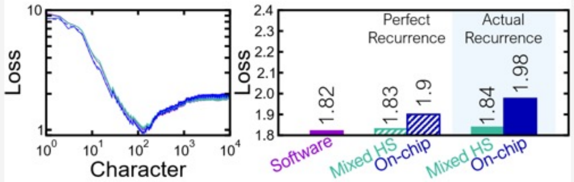
ResNet-9 layers on tile for CIFAR-10 classification
(All MVMs done on-chip. Pooling in software.)



Classification Accuracy = 85.6%
Software Baseline = 88.4%

1 Layer LSTM (RNN)
Alice-in-wonderland
Character Prediction





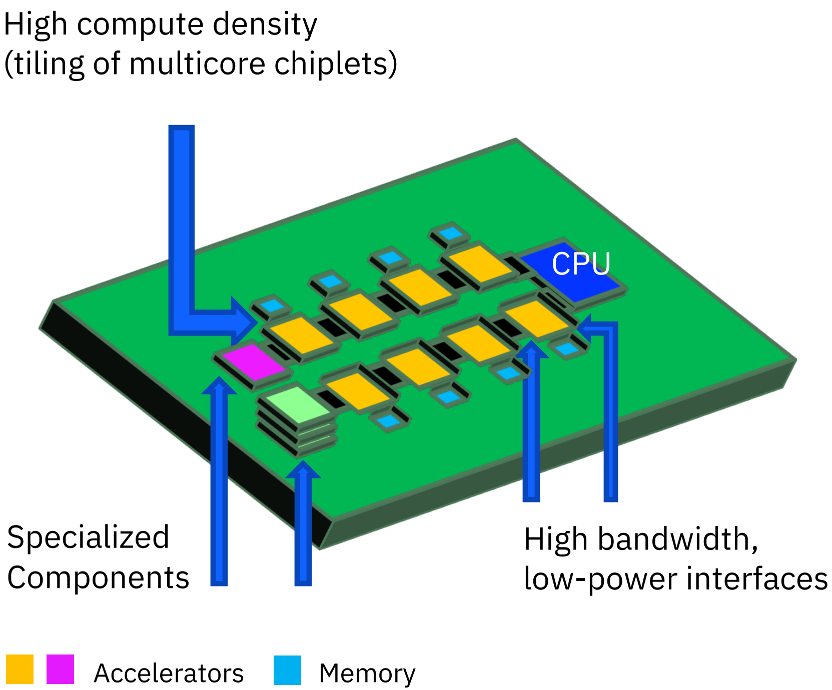
Heterogeneous Integration platform for AI

What's needed:

Interfaces between components

- High bandwidth (Gbps/mm)
- Energy-efficient (pJ/bit)
- Area-efficient (Gbps/mm^2)
- Standards to allow connectivity between wide variety of components

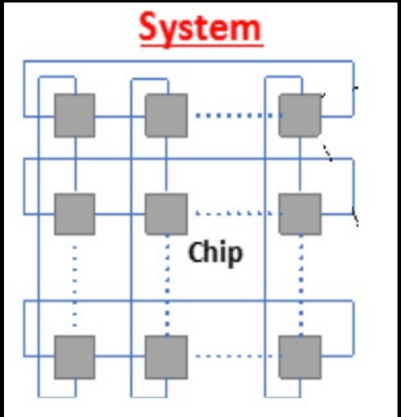
Heterogeneous Integration Technologies

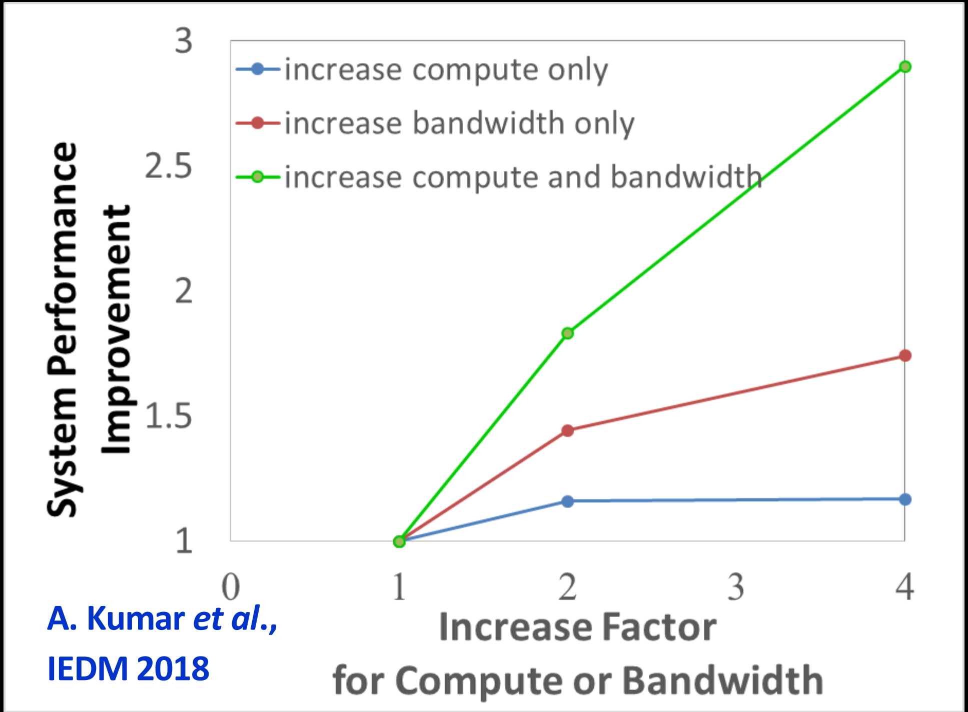


© 2021 IBM Corporation

Memory requirements

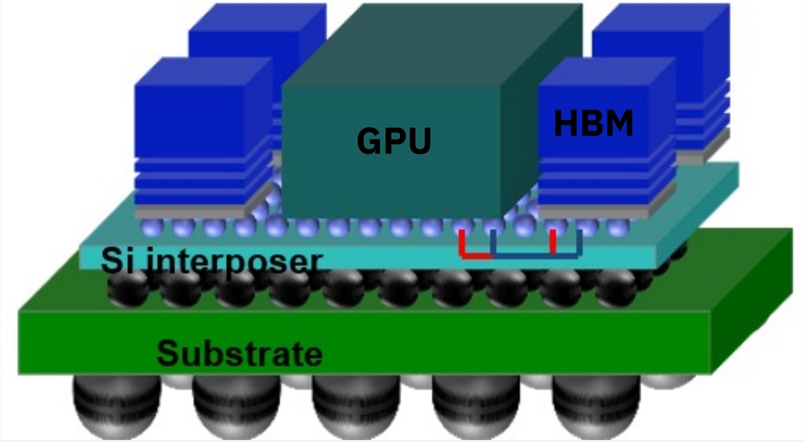
Multi-chip network of AI accelerators training Resnet-50
(Each chip has several AI cores from: B. Fleishcher *et al.*, VLSI 2018)

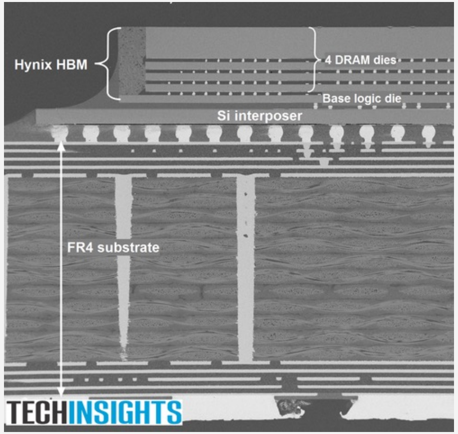




Memory bandwidth increase gives best “bang for buck”

Today’s GPUs





Key Attributes:

- Compute and memory closely coupled
- Si Interposer provides interconnect density
  - C4 scaling
  - Tight pitch wiring groundrules
- Utilizes standard organic substrate technology

Limitations:

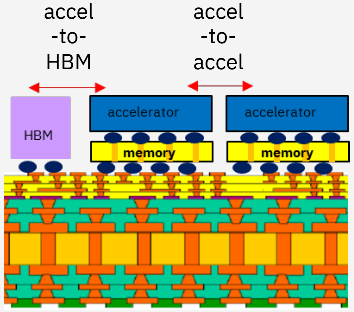
- Si Interposer body size
- Insertion losses associated w/ Si Interposer
- Cost (Si fab processing)
- Closed ecosystem

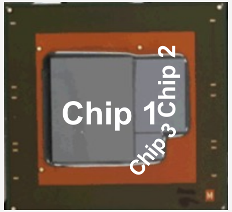
Our HI focus areas

Increasing complexity / time to market

HDI Laminate

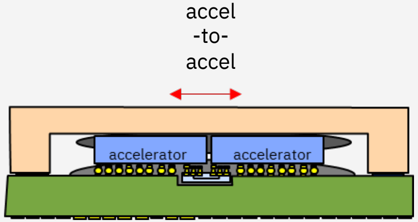
Enables tight pitch die interconnects at lower cost

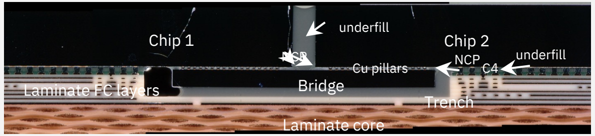




Si Bridge (DBHi)

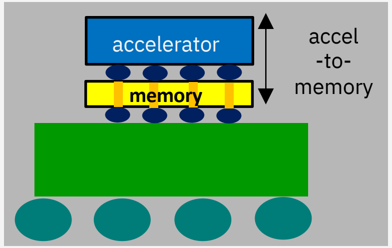
Higher connectivity, flexible configuration

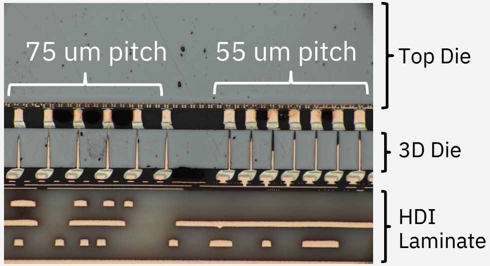




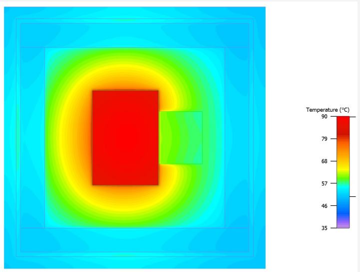
3D Integration

Highest interconnect density, scalable





Simulation & Modeling



End user AI testbed

End-to-end environment for learning, development, test & simulation of AI leveraging IBM’s state-of-the-art AI software tools and innovations

AI Supercomputer AiMOS

High-performance AI Supercomputer with a mix of commercial and pre-commercial tools

IBM Public Cloud

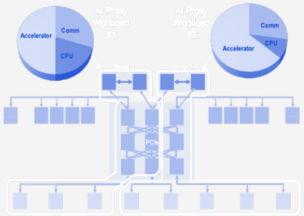
Use a consumable suite of common Data Science tools

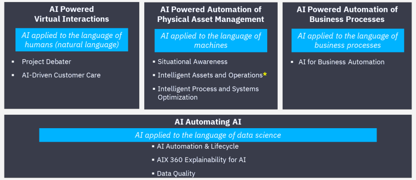
Composable Testbed

Experiment with various system-level topologies and configurations

AI Research Software Toolkits

State of the art AI research Innovations for AI-powered automation





AI Supercomputer powering key COVID-19 Research

HPC COVID-19 Consortium

Cleveland Clinic

Multi-Epitope Vaccine Design. *"Repurposing of FDA-Approved Toremifene to treat COVID-19 by blocking the spike glycoprotein and NSP14 of SARS-CoV-2. All simulations were done in AiMOS using GROMACS 2020",* Dr. William R. Martin and Dr. Feixiong Cheng

>48352 node-hours and over 6078 jobs in just Q2 2021 in AIMOS

Stony Brook University

Intelligent Platelet Dynamics. *"We have developed AI/machine learning algorithms to extract the basic platelet geometrical data to understand the mechanisms of blood clot formation",* Dr. Yuefan Deng

> 500x speedup with CPU-GPU complexes > 97% accuracy in platelet dynamics & mechanics

>13,413 node-hours and 221 jobs in Q2/2021 in AiMOS

Weill Cornell Medicine

Simulations of molecular mechanisms of SARS-CoV-2 interactions with membranes to enable the design of small molecule inhibitors of viral entry. *"Development of the first atomistic model of the fusion peptide region of the viral spike protein, and the first large scale molecular dynamics simulations of the membrane penetration process by this region that informed subsequent AI/ML-enhanced protocols for discovery of inhibitors of this first step in the process of infectivity, were all carried out on the AiMOS computer ",* Dr. Harel Weinstein and Dr. George Khelashvili

>48,937 node-hours and 17,047 jobs in Q2 of 2021 in AiMOS

IBM Research AI Hardware Center / © 2021 IBM Corporation

Open-source resources to evaluate analog AI technologies

Analog Hardware Toolkit

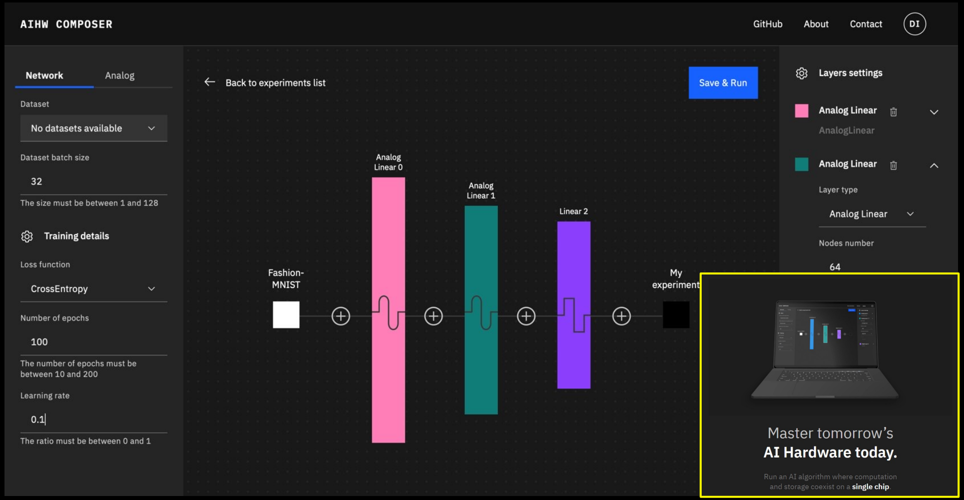
https://github.com/IBM/aihwkit

- ❖ Open-Source python toolkit for exploring in-memory computing devices for AI (deep learning) together with systems pillar
- ❖ Integrated with Pytorch
- ❖ Analog NN modules (fully connected layer, convolutional layer)
- ❖ **Explore Analog DNN training** using analog mat-vec and rank-1 update along with analog-specific SGD optimizers
- ❖ **Explore Analog DNN inference** with drift and statistical noise models
- ❖ Ready to download and install (using **pip**):
- ❖ `pip install aihwkit`

Analog Composer

https://aihw-composer.draco.res.ibm.com

- ❖ Web interface for exploration of Analog AI technology for DL training
- ❖ Explore performance of various NVM devices, models & training algorithms



Thank you

ibm.co/ai-hardware-center