EnCharge AI

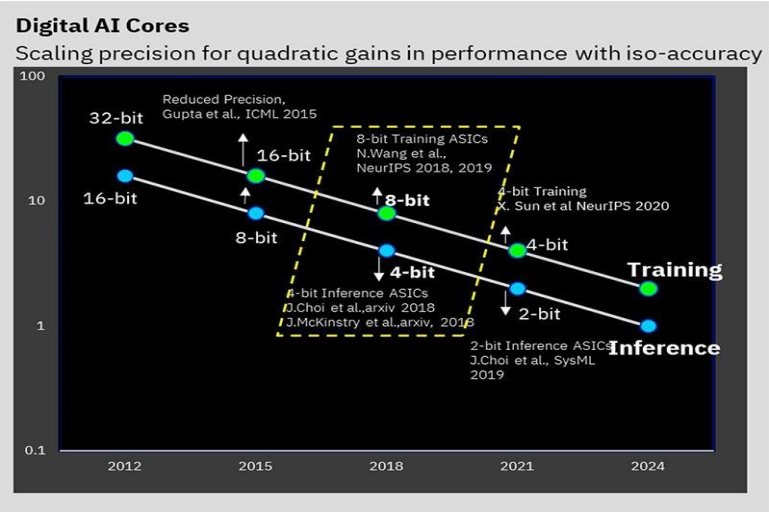
The 1000X Play for 2035?

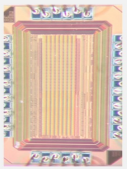
Kailash Gopalakrishnan

CTO, EnCharge AI

NSF Workshop on AI + HW 2035

Retrospective: Past 10 Years in AI Compute





130 nm
[VLSI'16]
[JSSC'17]

65nm
[VLSI'18]
[JSSC'19]

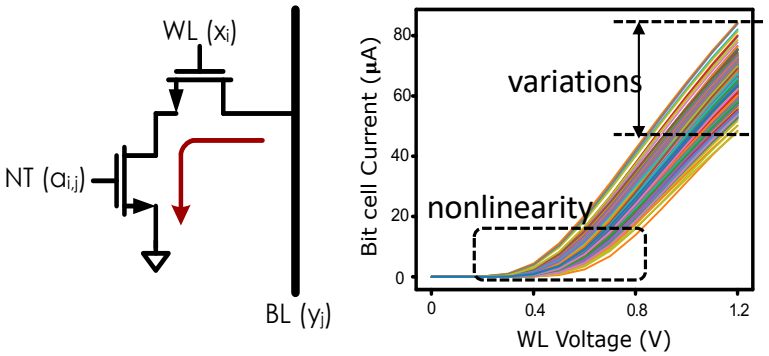
- **Precision Scaling Roadmap**
  - Leveraged quadratic benefits in throughput / efficiency / performance with precision
- **Down to 8b/4b (and even 2b/1.5b) today**
  - Not much room left for scaling (except in specific niche cases / models)
- **Silicon scaling below 2 nm is also incredibly challenging**
  - Limited to no improvements in density, cost, performance or energy
- **Early work on In-Memory Computing was seeded (2015)**
- **What work should we seed today so that 2035 continues to thrive in terms of AI-HW innovations?**

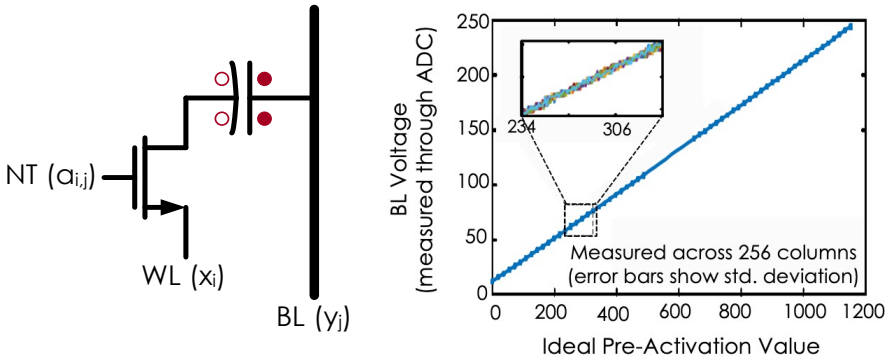
https://research.ibm.com/blog/ai-chip-precision-scaling

Today: Analog In-Memory Computing (Options)

EnCharge AI

- **Current accumulation via transistors**
- **Charge accumulation via capacitors**





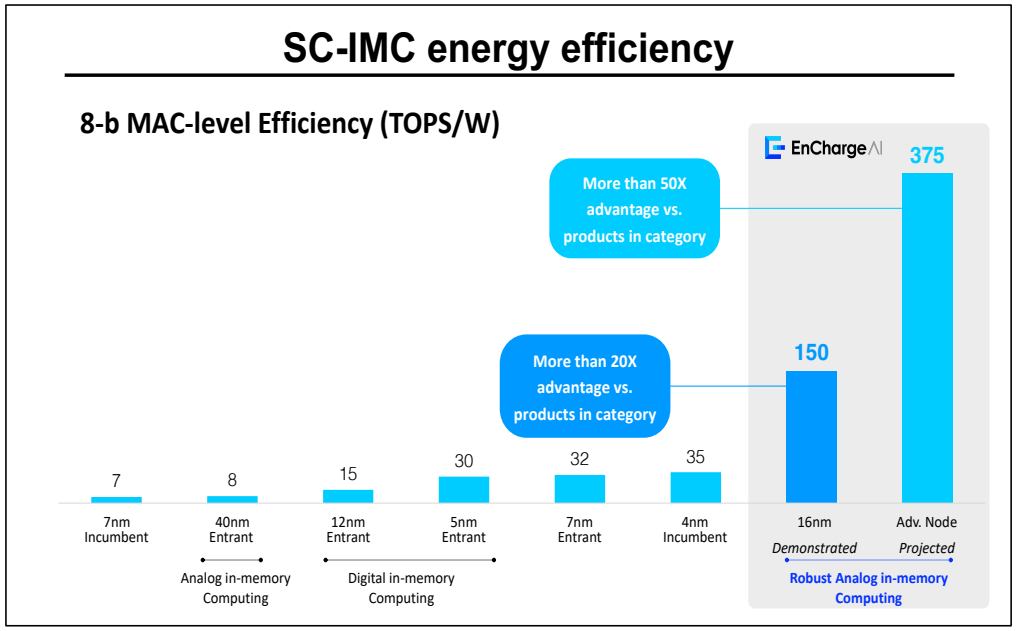
Harmed by process variability and noise	Requires costly, niche fab options	Can't support most-advanced AI models

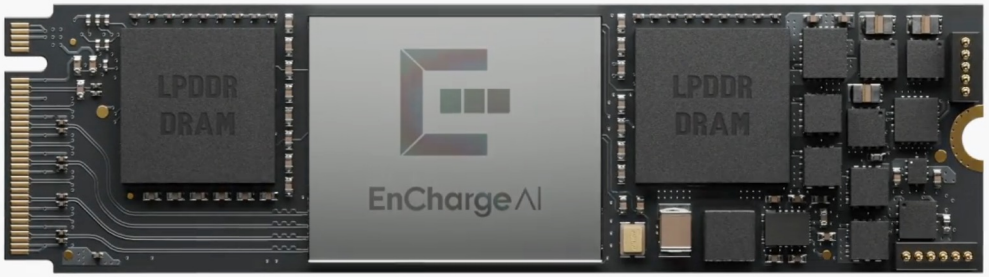
Examples: PCM, RRAM, Flash-based (Mythic),…

Intrinsically-precise metal capacitors	No special process options, scales like digital	Broad, accurate AI model support

Core EnCharge technology

EnCharge IMC Technology Attributes





Focus on AI-PC, Physical-AI Inference: High efficiency

- **Very high density & power-efficiency**
  - 10-20X better metrics than digital (including in Digital-IMC)
  - Optimized for compute (not for storage)
- **Accuracy identical to 8b Digital**
  - Model accuracy is preserved for state-of-the-art models
- **Scalability to aggressive Si nodes**
- **NOT WEIGHT STATIONARY**
  - Supports hardware multiplexing
  - Looks and feels like a digital tensor-core
- **Programmability & architectural scalability** - to support a wide range of models

The Next 10 Years?

EnCharge AI

- **Need to tackle all 3 bottlenecks: Compute, Communication & Memory**
  - Architectural innovations to minimize other bottlenecks within the chip
  - Extending Analog Computing to Training
  - Minimizing memory energy – tighter packaging with memory (3D-stacking?)
  - Enabling efficient data communication paradigms (Chiplets, Co-packaged optics…) to get to < 1 pJ/bit
- **Application / AI Model Evolution**
  - Autoregressive -> Diffusion Models (compute-bound vs. memory-bound)
  - HW substrate needs to be programmable enough to support evolution in model architectures