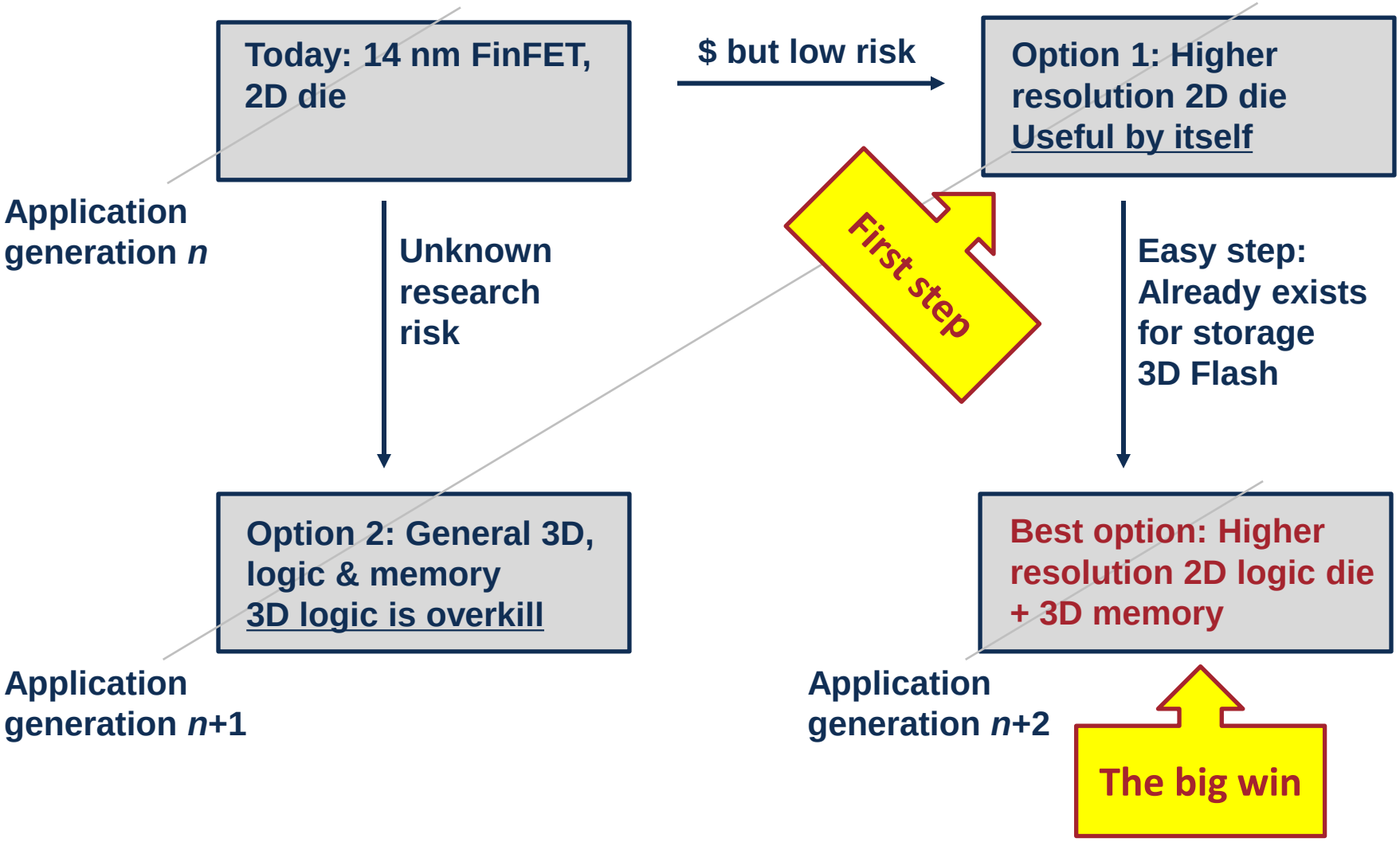
High Lithographic Resolution & 3D:
the 3D memory is the important part

Erik P. DeBenedictis
Center for Computing Research, Sandia
IEUVI: 3D-EUV Strategic Discussion
September 10, 2017

Approved for Unclassified Unlimited Release
tracking number SAND2017-9983 PE

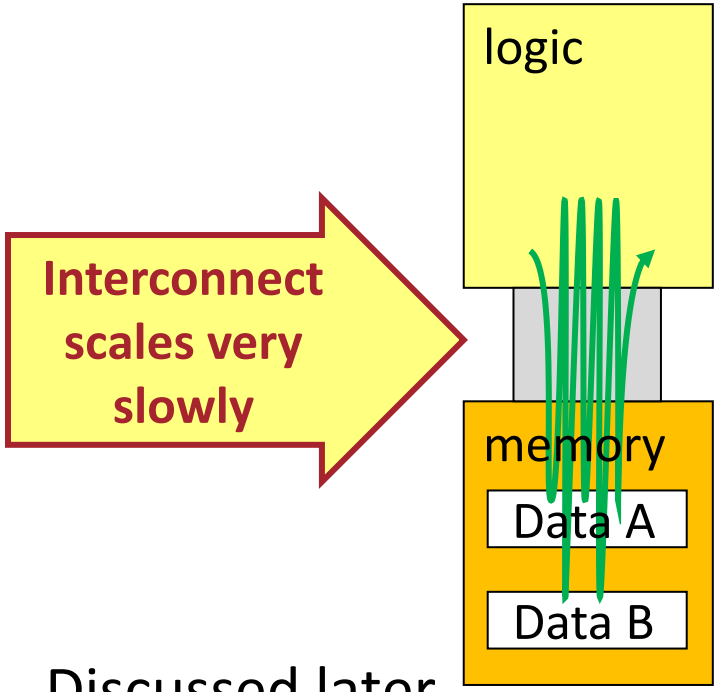
Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

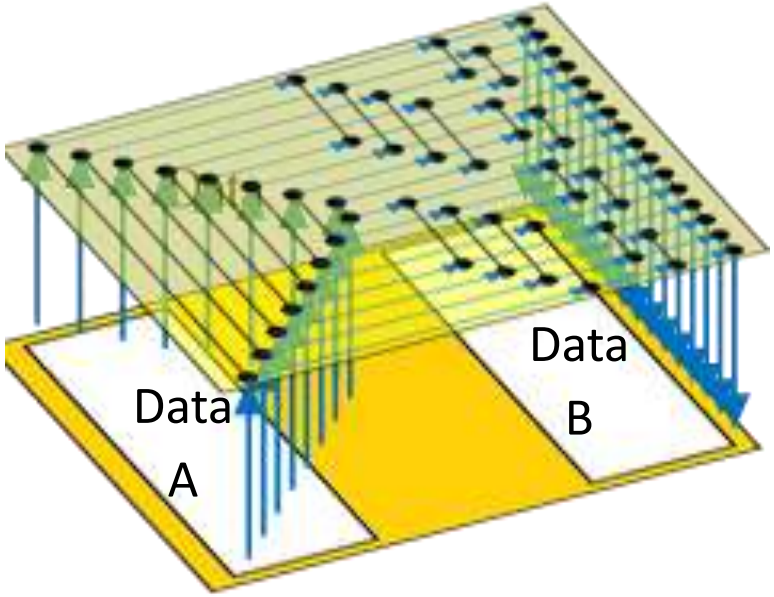
Two application generations are on the table



Data modality

- 3D relieves the penalty for information to switch between logic and memory
- Changes algorithms





- Discussed later

3D advantage example: sorting algorithms

Sandia National Laboratories

Sorting algorithms

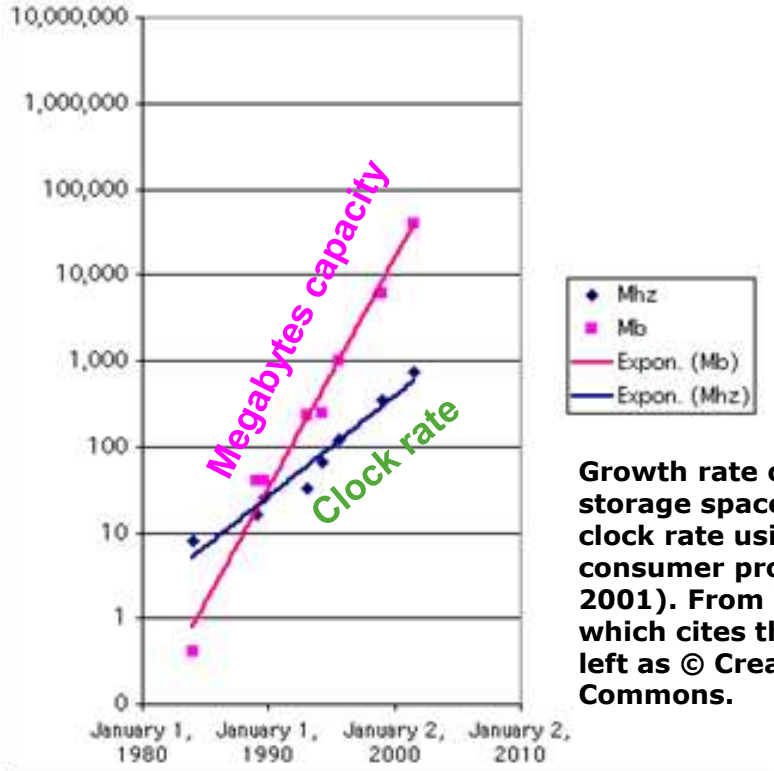
- Sorting network $O(\log^2 n)$, but requires parallelism
- Quicksort $O(n \log n)$ runtime based on von Neumann processor operation count
- Sorting network $O(n \log^2 n)$ based on von Neumann processor operation count

Result

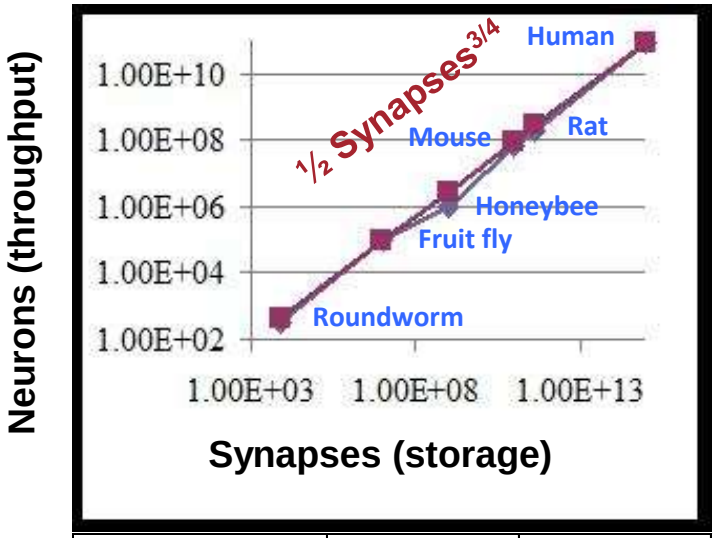
- With separate logic and memory chips
  - Quicksort $O(n \log n)$ fastest
  - Sorting network $O(n \log^2 n)$ slower
- With integrated logic and memory
  - Sorting network $O(\log^2 n)$ fastest
  - Quicksort $O(n \log n)$ slower

3D Scaling and Applications I

- Applications require more memory as Moore's law proceeds



Growth rate of HDD storage space compared to clock rate using Apple consumer products (1984-2001). From Wikipedia, which cites the diagram to left as © Creative Commons.



	Synapses	Neurons
Roundworm	7.50E+03	3.02E+02
Fruit fly	1.00E+07	1.00E+05
Honeybee	1.00E+09	9.60E+05
Mouse	1.00E+11	7.10E+07
Rat	4.48E+11	2.00E+08
Human	1.00E+15	8.60E+10

Source: Wikipedia

3D Scaling and Applications II

Why?

- Fruit fly heads towards food, avoids hands trying to kill it
- Humans use much more information to navigate
  - Who owns the property you're on?
  - Don't run out of gas
  - Take shortest path
  - Go through rain but avoid pedestrians
- A lot of new apps are in the AI domain
  - Which model appropriate for a self-driving car?
  - (Don't drive the Tesla through the semi with a landscape painted on the trailer)

3D Scaling and Applications III

Scaling at constant clock rate

- 2D chip will scale until logic+SRAM fills the reticle
- 2D logic + 3D memory will scale further. Until:
  - Logic fills the reticle OR
  - 3D memory reaches maximum layers OR
  - Heat dissipation limit reached

- General 3D: 3D logic + 3D memory will scale until:
  - Heat dissipation limit reached, given lower quality 3D transistors

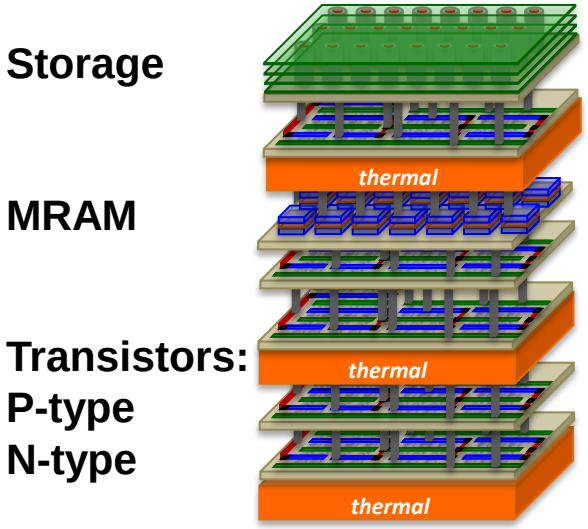
3D Memory easier than 3D logic

Memory only

- Regular structure
  - Ideally 1 litho step for *n* layers
  - Amenable to error correction
- 3D flash in production
  - 48, 72, … layers
  - Available in stores and online

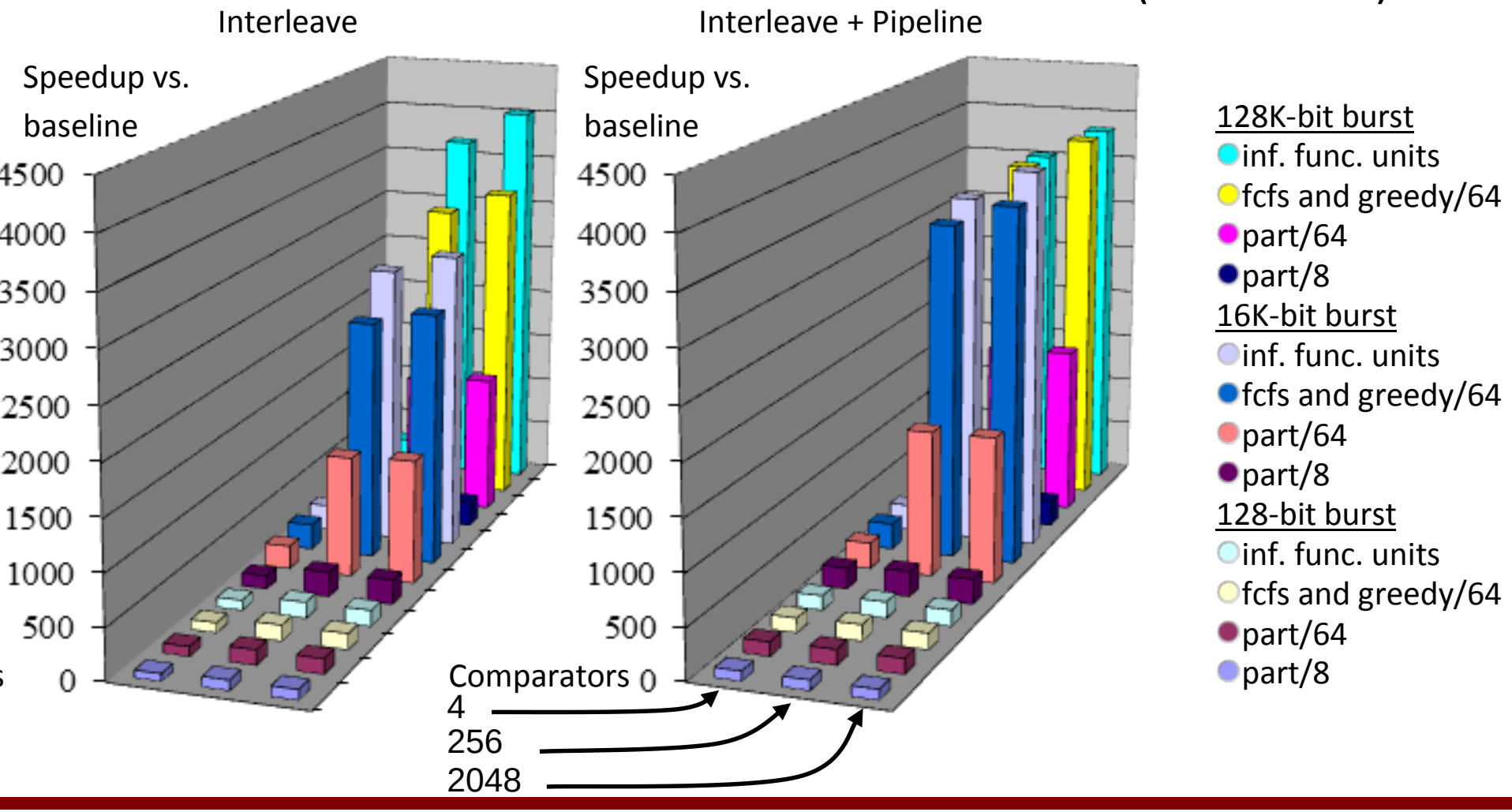
General 3D (N3XT)

- Arbitrary design
  - Many litho steps increase cost and reduce yield
  - Error tolerance much more difficult to achieve



Does it matter? Test case: Superstrider

- Simulations show 1000 × benefit from 3D alone (sometimes)



Assumptions, or what could go wrong

- We assume the 2D die can be filled with logic
  - Today's microprocessors tend to have a lot of memory, which is arguably to reduce heat flux
  - Can we reduce energy consumption per gate in the future?
  - Will architecture changes reduce power through reduced duty cycle?
- We assume a 3D memory as efficient as flash storage
  - Monolithic 3D Flash uses one litho step for all layers, however the process does not generalize to other memory types
  - Scientists assure me that an equivalent process will be available for the device that they are advocating (e. g. memristor), but I'm not convinced

Conclusions

- Should be able to continue Moore's law with:
  - 3D memory (monolithic, i. e. tighter *z*-direction coupling) +
    new algorithms +
    new or augmented architectures +
    3D RAM equivalent (where needed, as opposed to non-volatile)
  - 2D logic + 3D memory ought to do ($N_{memory} \approx FLOPS^{3/2}$)
- Can be done in multiple steps
  - Scale 2D (EUV) with logic+SRAM filling the chip
  - Move memory into third dimension and scale further
- Computer engineers will have their hands full with work on architectures and algorithms
- Other uses of 3D are good too, but not as potent

Thank you

See my Rebooting Computing Column in IEEE Computer

- April ‘16 Boolean Logic Tax
- June ’16 Learning Machines
- August ’16 Search for Secretariat
- October ’16: Help Wanted Turing
- December ’16 (see first page) →
- February ’17 Redefine Moore’s Law
- April ’17 Architecture’s Role
- June ‘17 Reversible Computing
- August ‘17 3D

REBOOTING COMPUTING

Computational Complexity and New Computing Approaches

Erik P. DeBenedictis, Sandia National Laboratories

Computational complexity analysis allows us to quantify energy-efficiency scaling potential—an important task for assessing research options.

As we search for new ways to increase both computer performance and energy efficiency, it would be helpful to be able to predict long-term potential in advance. Here, I'll show how computational complexity theory can quantify the energy efficiency potential of analog computing. The method could be applied to other computing approaches.

Analog computers are one option to restore growth in the computer industry. Such growth requires families of computers that can solve problems more cost-effectively over time, which today means improving energy efficiency. The improvement rate for the energy efficiency of digital computers has slowed, raising the question of whether analog computers could overtake them.

If analog and digital are viewed broadly as alternative computer implementations, then they should be subject to the same general principles. However, a specific digital computer's effectiveness depends on its architecture and the algorithms running on it. These correspond to the circuitry of an analog computer.

Here, I'll analyze digital and analog "neuromorphic" calculations using a computational complexity theory first developed for digital computer algorithms. The analysis doesn't find a winner but provides new insights into which approach has more potential.

COMPARING A COMMON FUNCTION

Meaningful comparison of analog and digital requires a computing task amenable to both approaches. I'll focus on artificial neural networks, where the comparison is between a digital implementation such as deep learning[2] and an analog neuromorphic implementation such as the atomic-wire circuit based on memristors.[3]

Biological neurons, which fill a role similar to N input logic gates, mathematically evaluate the computational primitive called dot or inner product. N "presynaptic" neurons generate signals that become inputs of the N input neuron under consideration. Each of the N signals v_i is multiplied by a synapse weight w_i and the products added to become the neuron's output. A digital implementation