Approved for unlimited release
SAND2017-5418C
Tracking Number: 623685

Artificial Intelligence, 3D, Architecture & "Superstrider"

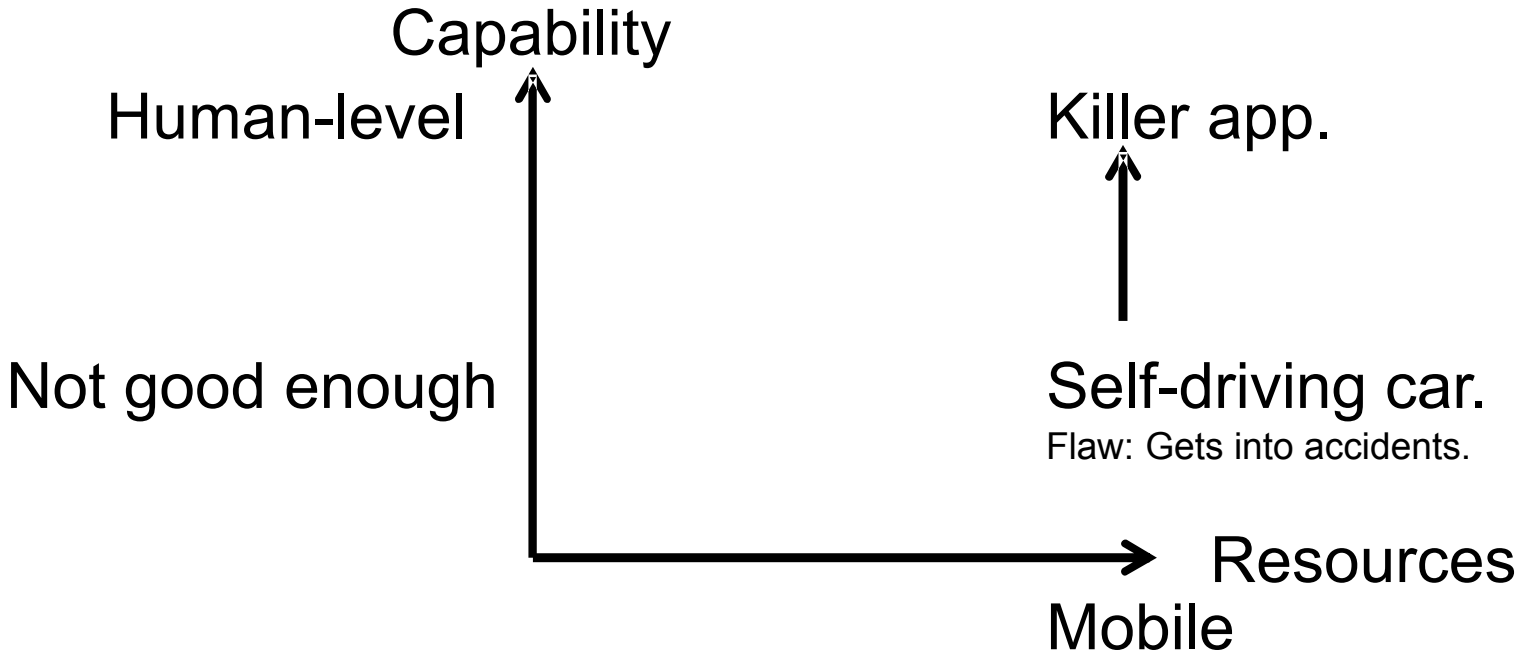
Erik P. DeBenedictis,[1] Jeanine Cook,[1]
Sriseshan Srikanth[2]
[1]Center for Computing Research, Sandia
[2]Georgia Tech

INC NANO May 10, 2017

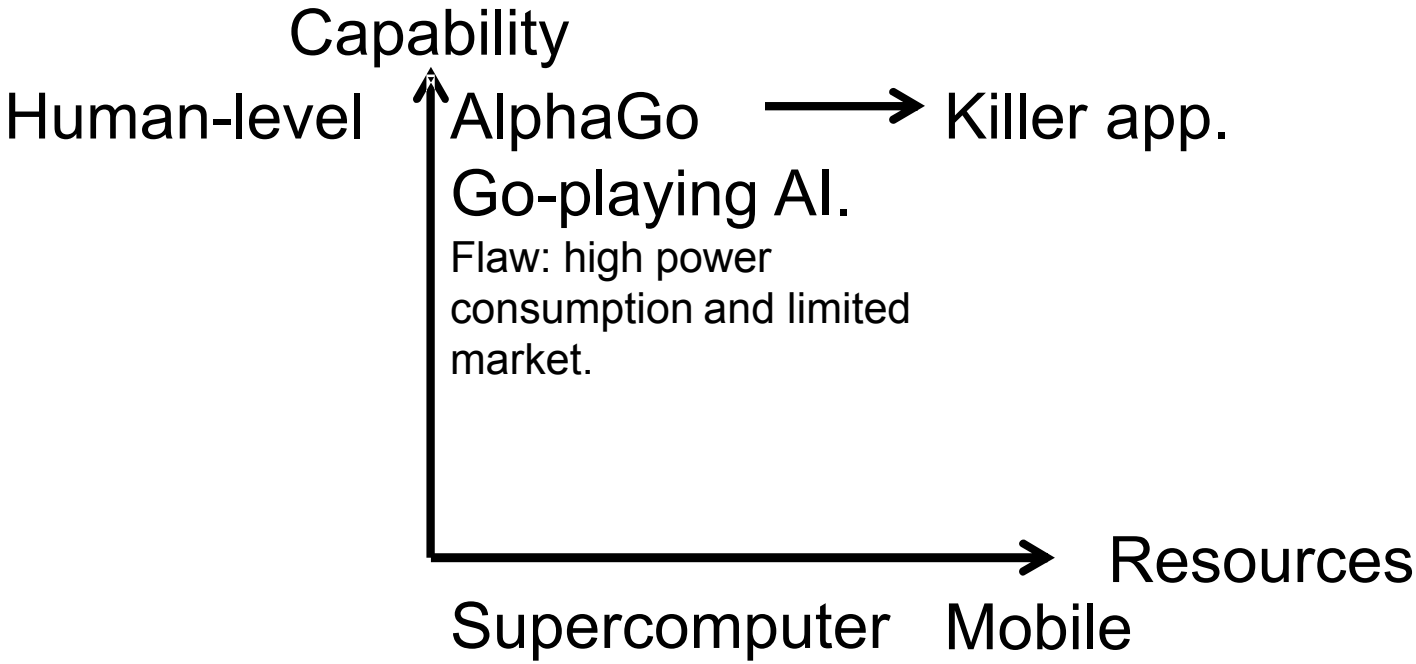
Paolo's Requests

1. Artificial Intelligence
   - How can we integrate AI into INCNANO/IRDS?
2. 3D and Architecture
   - Paolo's slide deck on ITRS is not smooth scaling, but discrete jumps or inflection points
     - High-K and FinFET
   - Can I tell the story of the next inflection point?
     - 3D and architecture

1. Rebooting Computing to Support Next “Killer App”



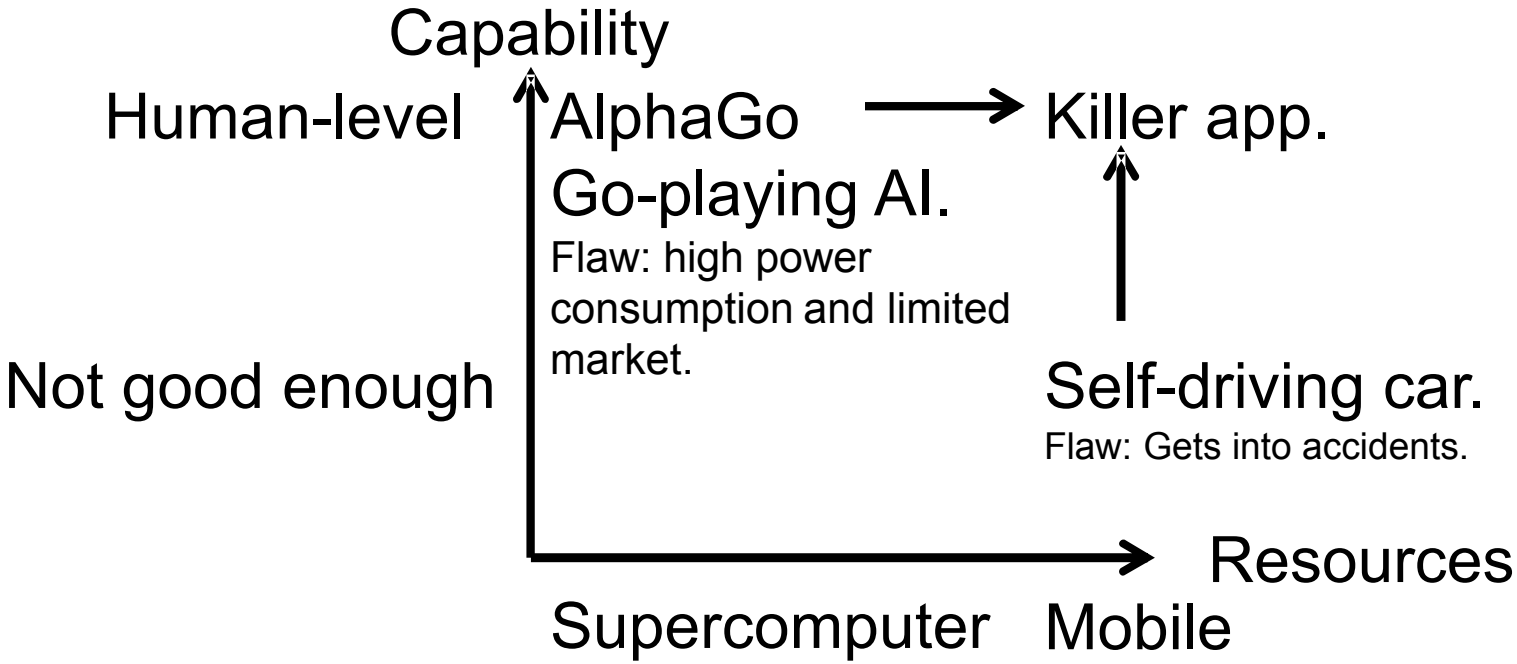
1. Rebooting Computing to Support Next “Killer App”



1. Rebooting Computing to Support Next "Killer App"

New "Killer Apps"

- We can program mobile killer app candidates on supercomputers



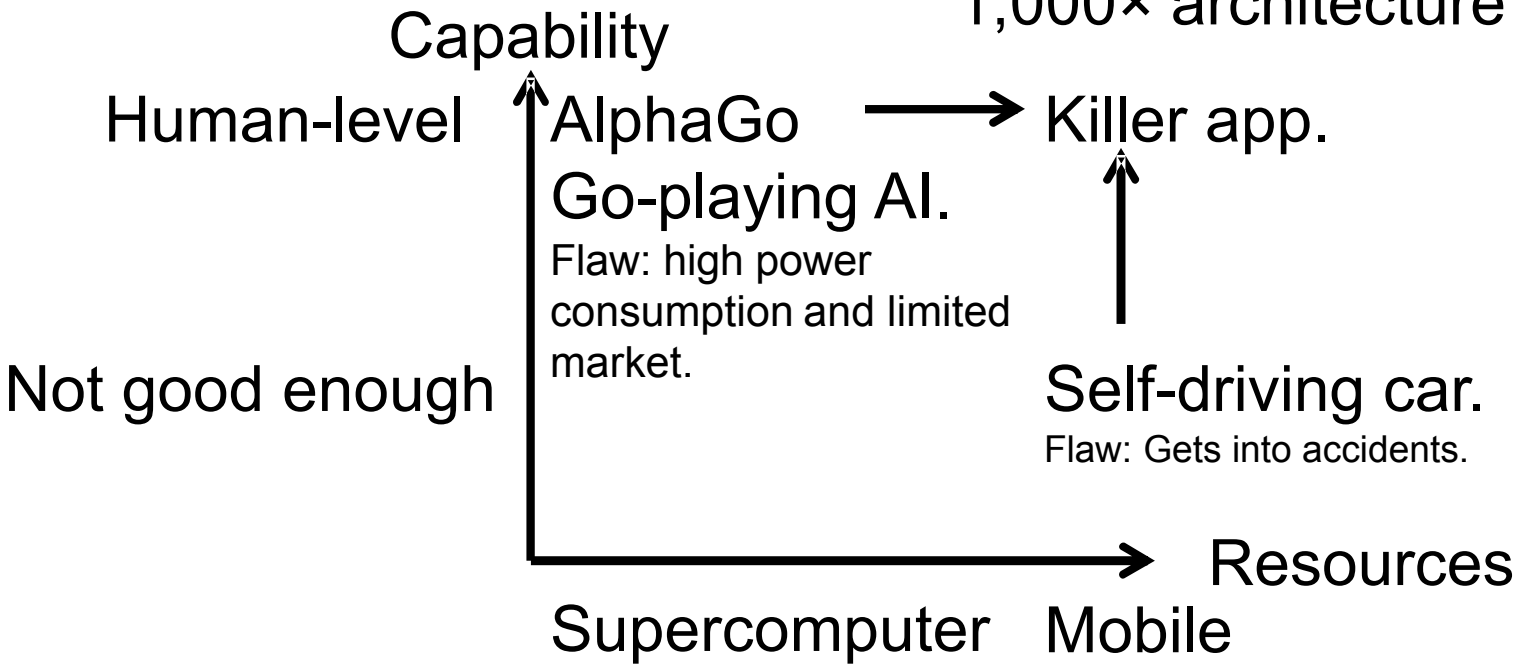
1. Rebooting Computing to Support Next "Killer App"

New "Killer Apps"

- We can program mobile killer app candidates on supercomputers

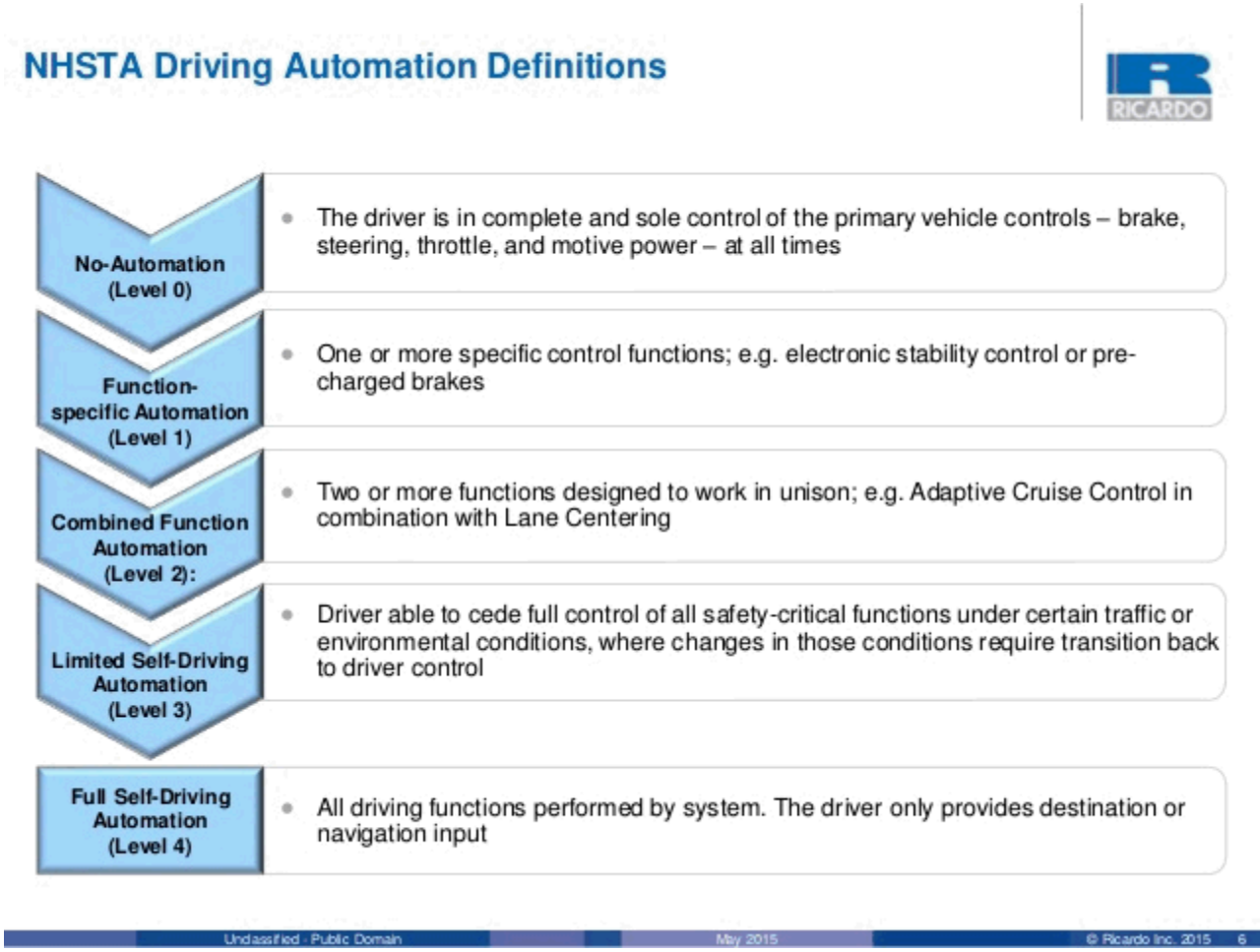
Applications pull

- Reduce a 100 kW cluster to a 1 W mobile.
- 10× devices, 10× 3D, 1,000× architecture



Backup: AI Roadmaps Based on Functional Milestones

- These are levels of application
- function



https://www.slideshare.net/AliMaleki2/cti-presentation-48709481
(claimed to be public domain)

Backup: But use Intense Technology

- Can we fuse functions with technology?

http://www.cnx-software.com/2015/12/07/renesas-r-car-h3-deca-core-processor-and-driverless-car-roadmap/
From Internet

Link Function to Technology

AI Milestone	Task Description	Technology/ Architecture	Power / Ops	Year
Full Self-drive	Drive car from sensors and visual cues	GPU	50 W / 10^y	20yy (safely)

Link Function to Technology

AI Milestone	Task Description	Technology/ Architecture	Power / Ops	Year
Driver assist	Identify cars in other land when driver activates turn signal	CPU	50 W / 10^x	2000-2009
Full Self-drive	Drive car from sensors and visual cues	GPU	50 W / 10^y	20yy (safely)
Fully autonomous mini-robot	Plan, move, and carry out missions	Neuro-morphic?	50 mW / 10^z	20zz

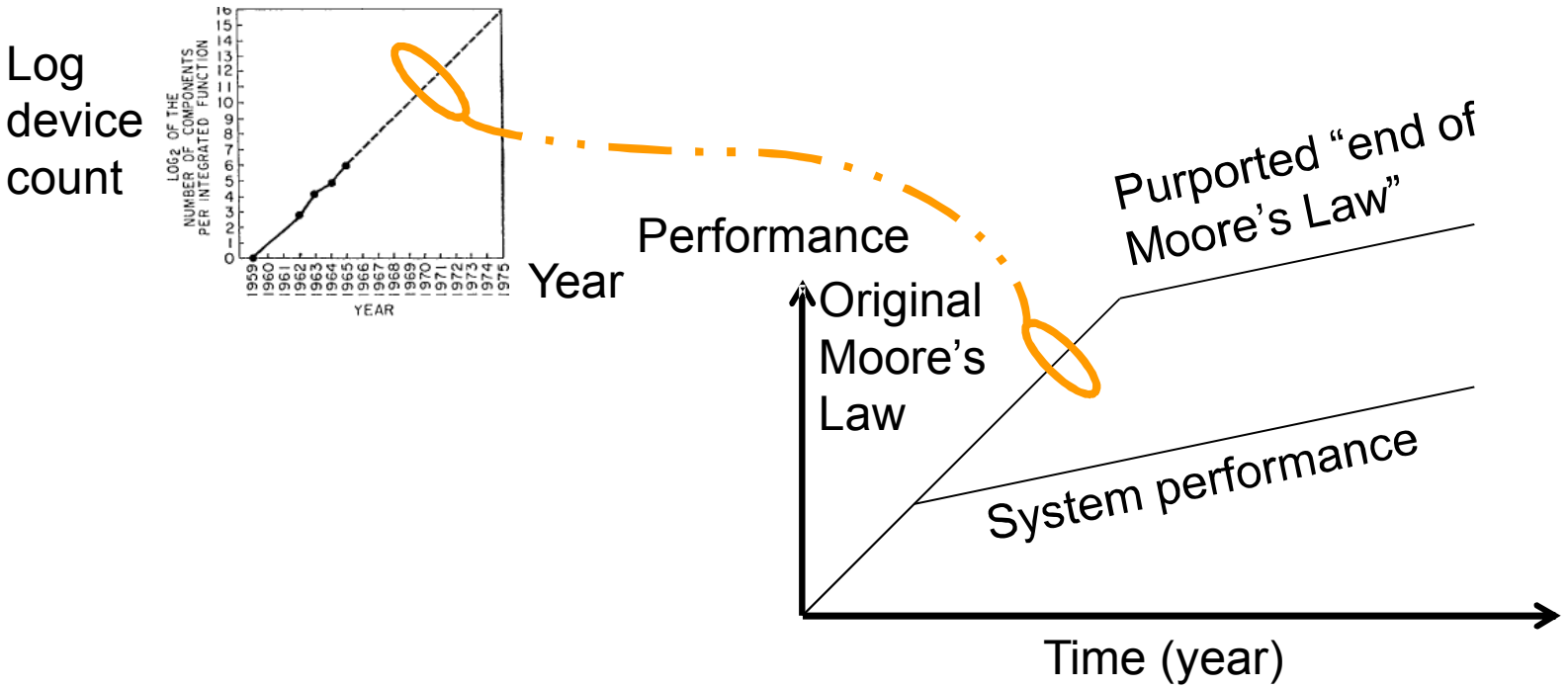
2. 3D and Architecture

Moore's Law

- Projection
  - I know Devices ≠ Performance
  - Figure 3 in Moore's article

What Happened

- Device performance didn't translate to system level



Backup: How Much Scaling Didn't Get to the System Level?

Binary Throughput

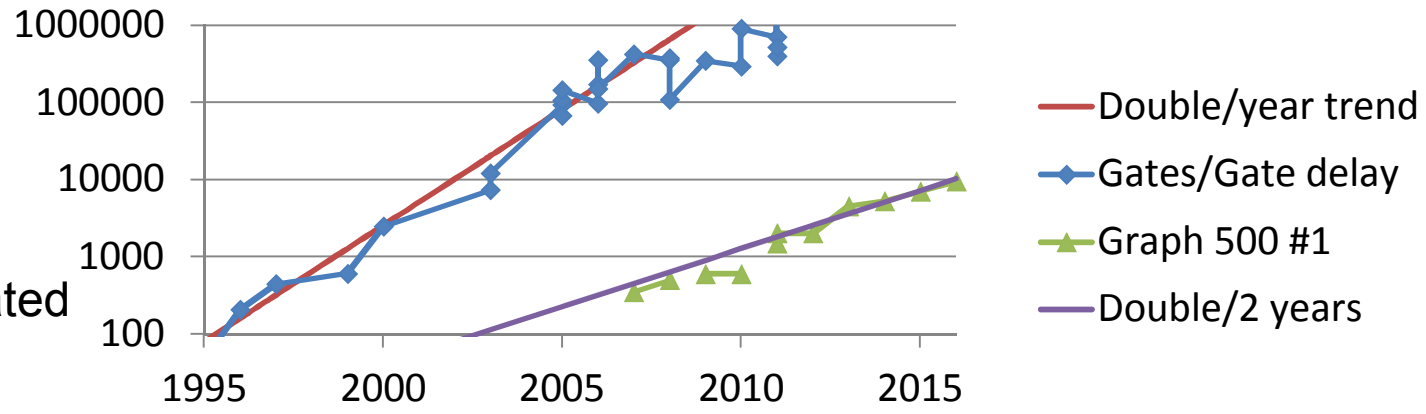
- BIT rate
  - Gates per cm^2 chip ×
  - Max rate/chip
- Doubled each year

Green 500

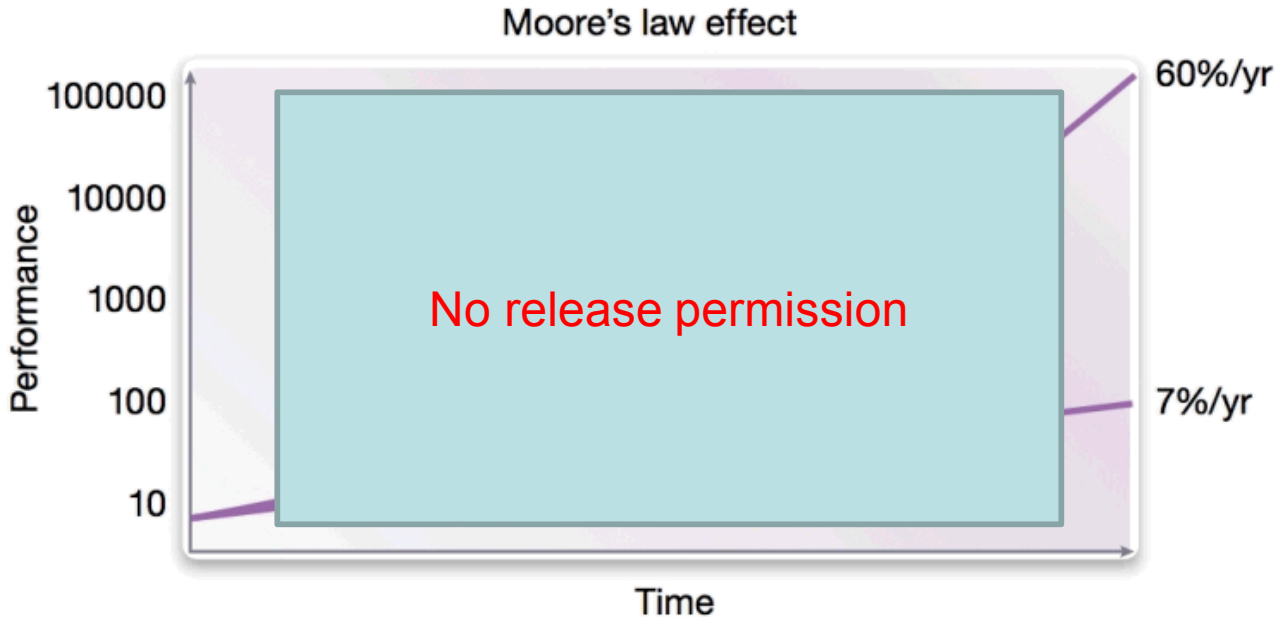
- Benchmark for energy efficient computing
- Rated in flops/watt
- Doubles each 2 years

Gap grew about √2 per year, for a long time!

Axes log scale but uncalibrated



Backup: Performance Gap



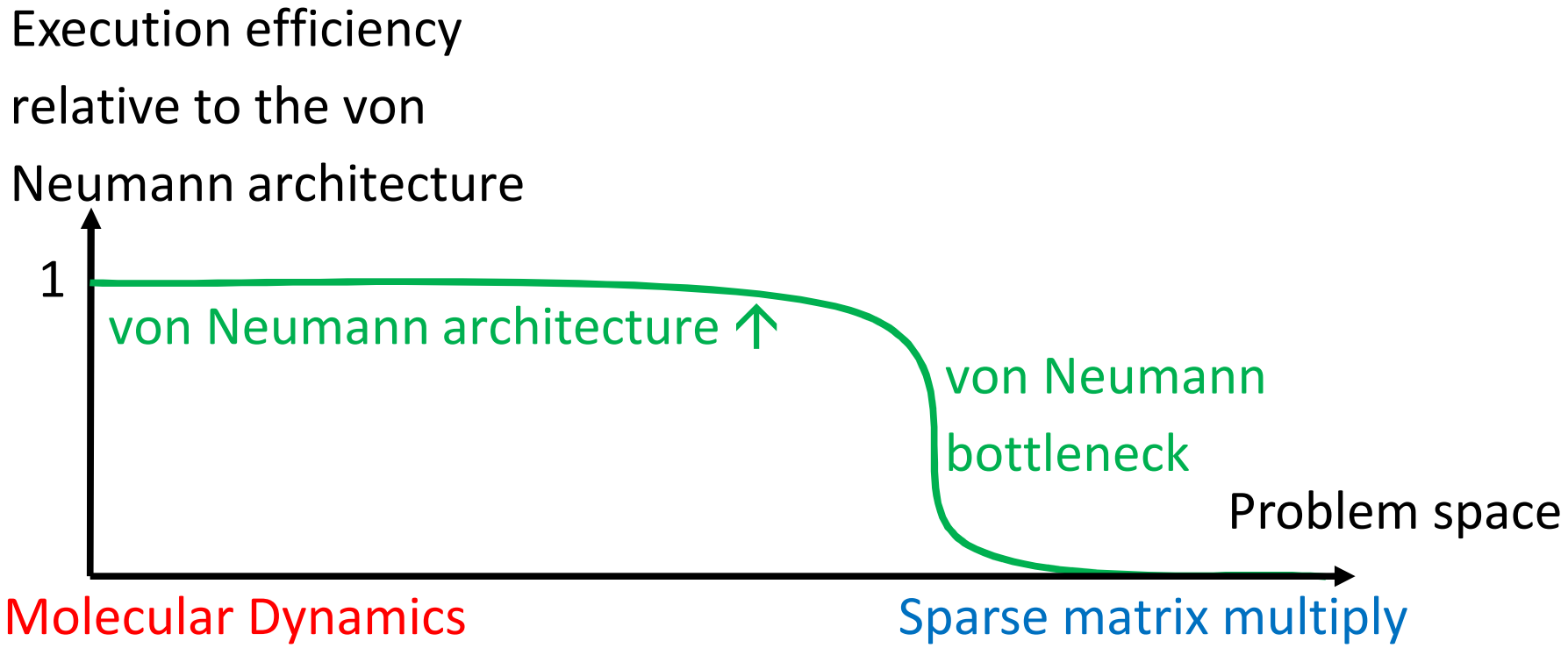
https://motherboard.vice.com/en_us/article/memory-is-holding-up-the-moores-law-progression-of-processing-power
Says "Source SYNOPSIS" ?

von Neumann Bottleneck

Execution efficiency relative to the von Neumann architecture

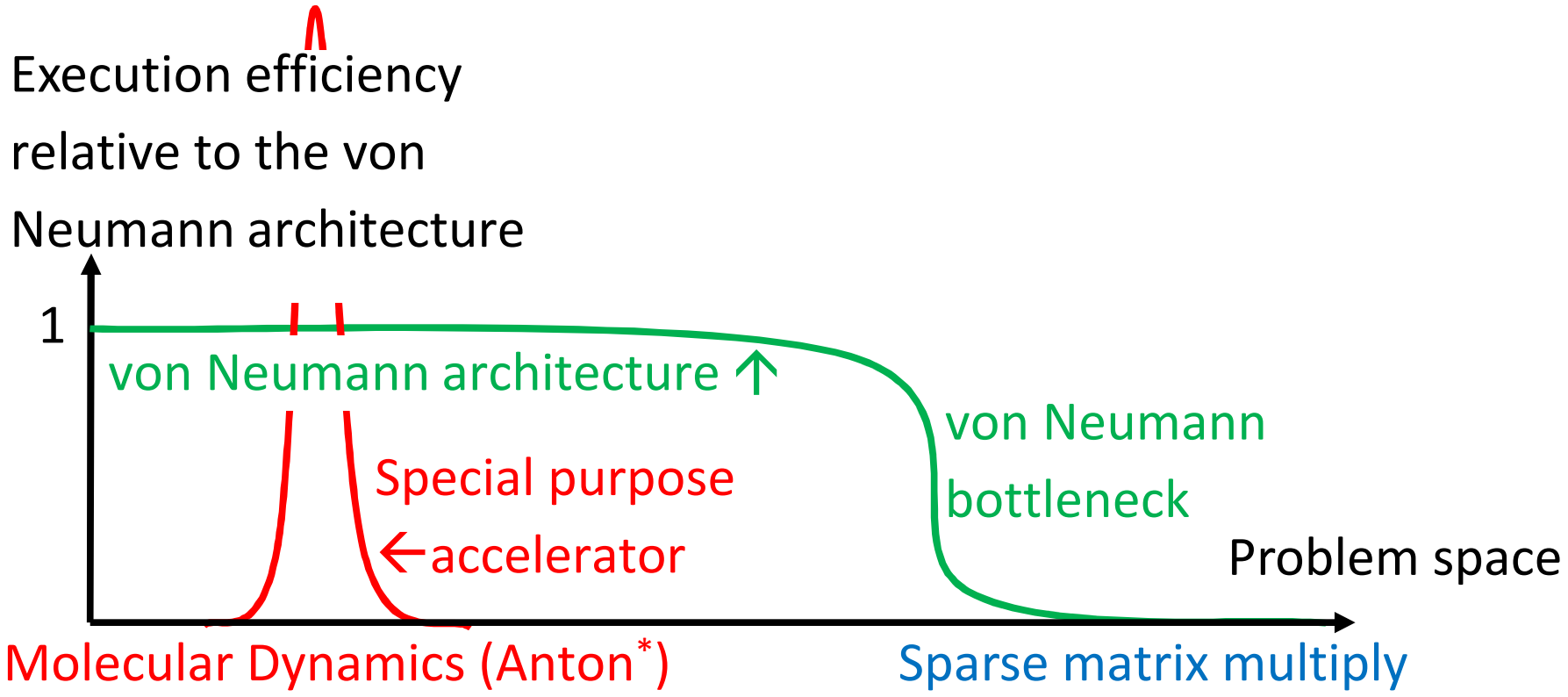
Definition of axes:

von Neumann Bottleneck



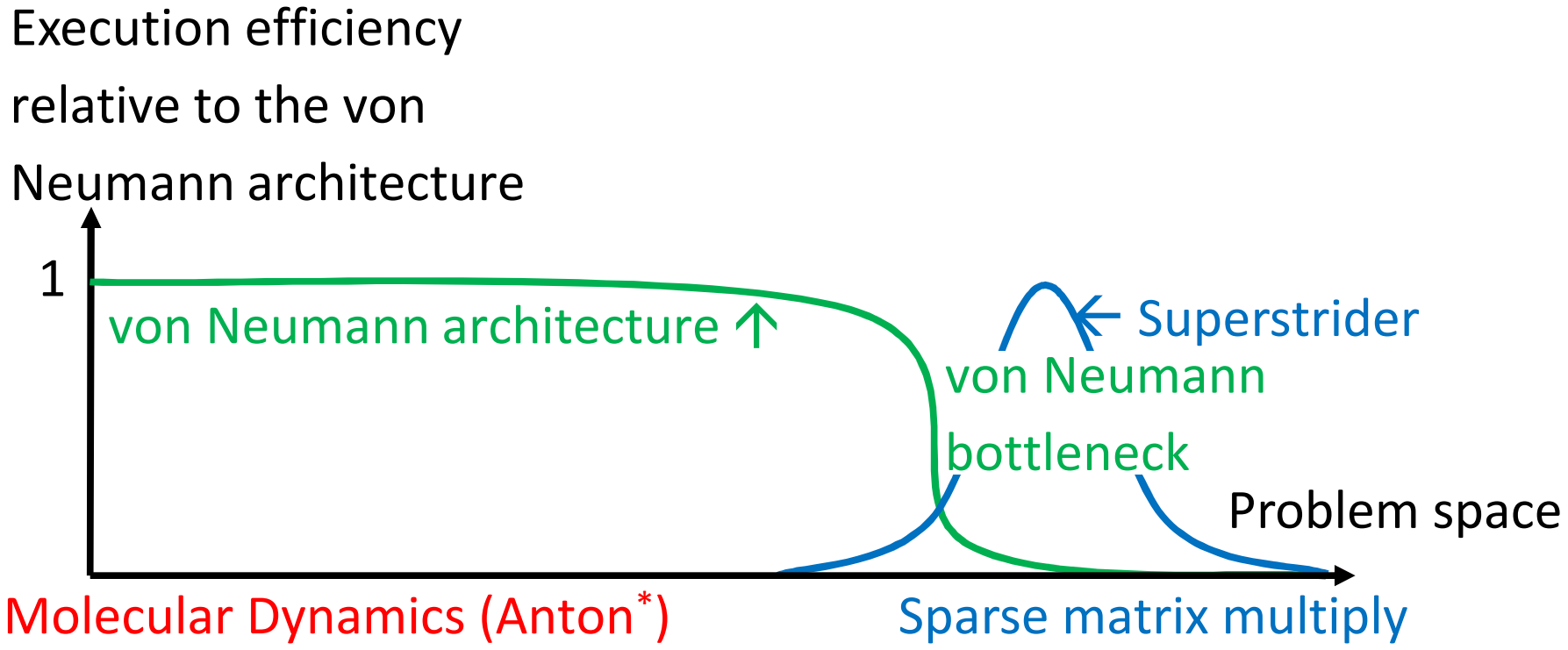
Definition of axes:

von Neumann Bottleneck



*Anton = Molecular dynamics ASIC by D. E. Shaw research

von Neumann Bottleneck

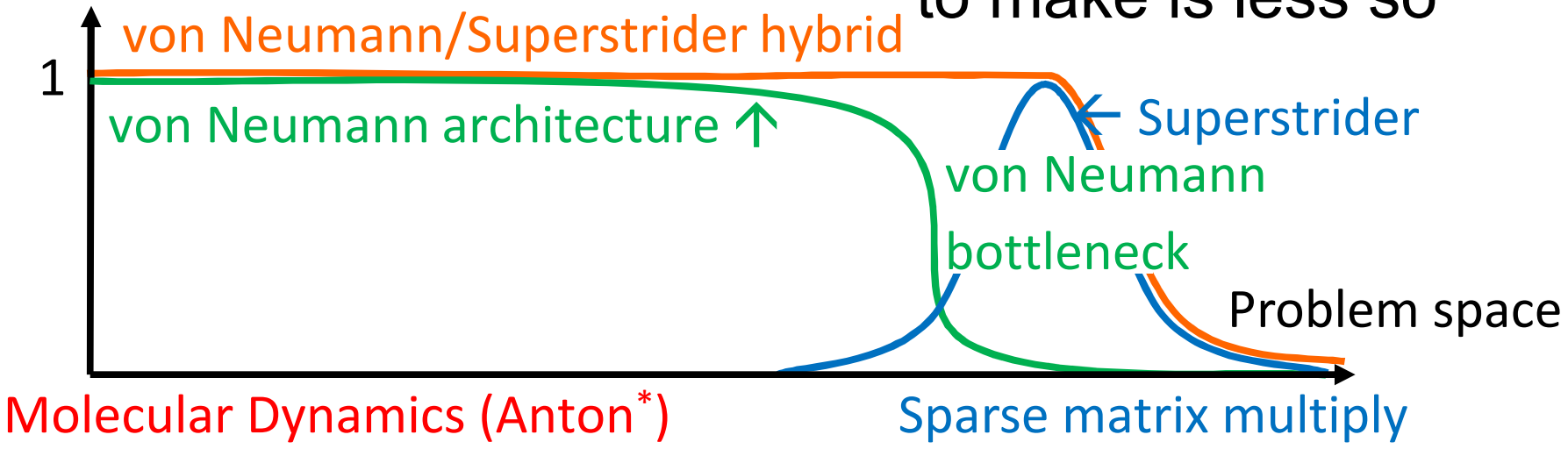


*Anton = Molecular dynamics ASIC by D. E. Shaw research

von Neumann Bottleneck

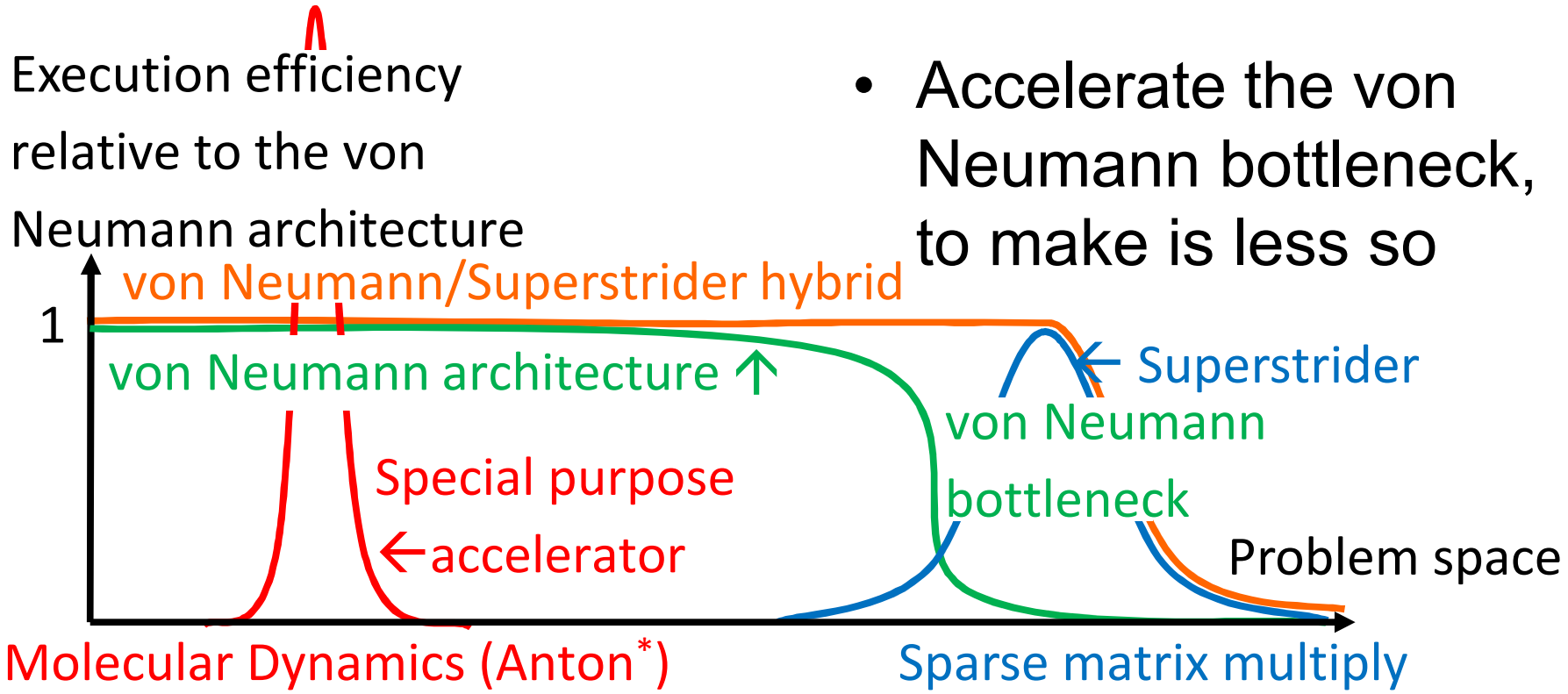
Execution efficiency relative to the von Neumann architecture

- Accelerate the von Neumann bottleneck, to make is less so



*Anton = Molecular dynamics ASIC by D. E. Shaw research

von Neumann Bottleneck



- Accelerate the von Neumann bottleneck, to make is less so

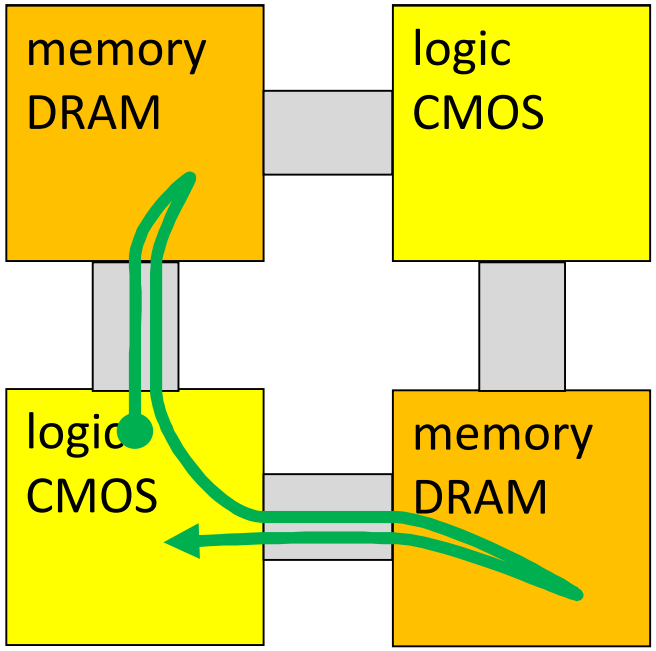
Definition of axes:

*Anton = Molecular dynamics ASIC by D. E. Shaw research

Underlying Source of the von Neumann Bottleneck and How To Fix It With 3D

2D/von Neumann

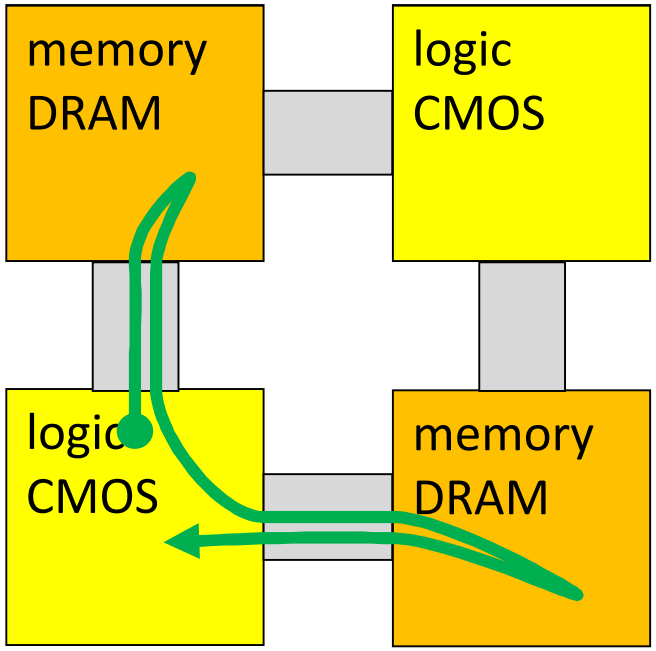
- Inefficient: repetition of { memory access from large pool + logic }



Underlying Source of the von Neumann Bottleneck and How To Fix It With 3D

2D/von Neumann

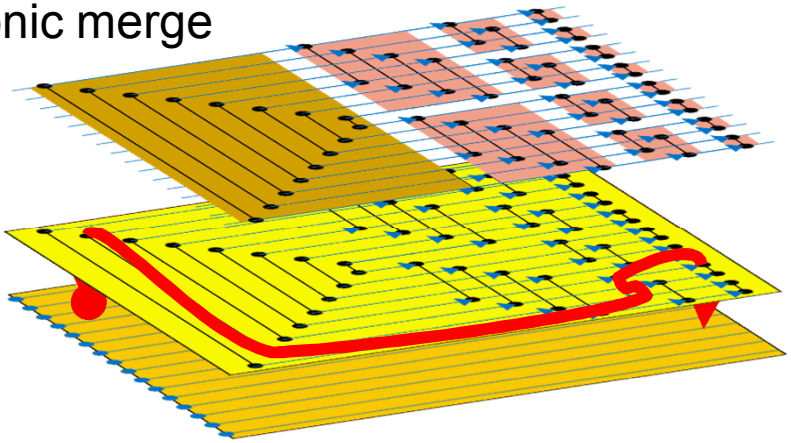
- Inefficient: repetition of { memory access from large pool + logic }



3D and New Architectures

- Micron-level layer shift reduces delay and energy
- Enables placement
- Some algorithms become more efficient

Data dependency diagram for Bitonic merge



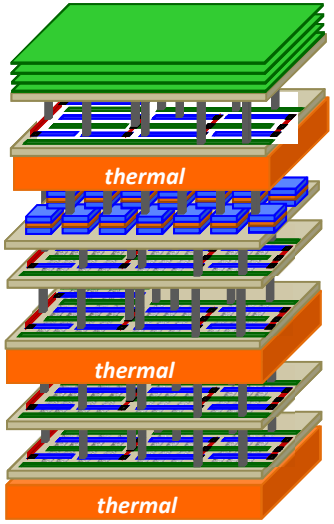
3. Superstrider Test Case Sandia/Ga Tech

Architecture

- An accelerator for the von Neumann bottleneck
  - To make it less of a bottleneck, not more
- Implements “associative arrays,” which are pretty general
  - Kepner, Jeremy, and John Gilbert, eds. *Graph algorithms in the language of linear algebra*. Society for Industrial and Applied Mathematics, 2011.

Simulation

- Simulated scaling scenario
  - 50× today with High Bandwidth Memory
  - 250× with N3XT

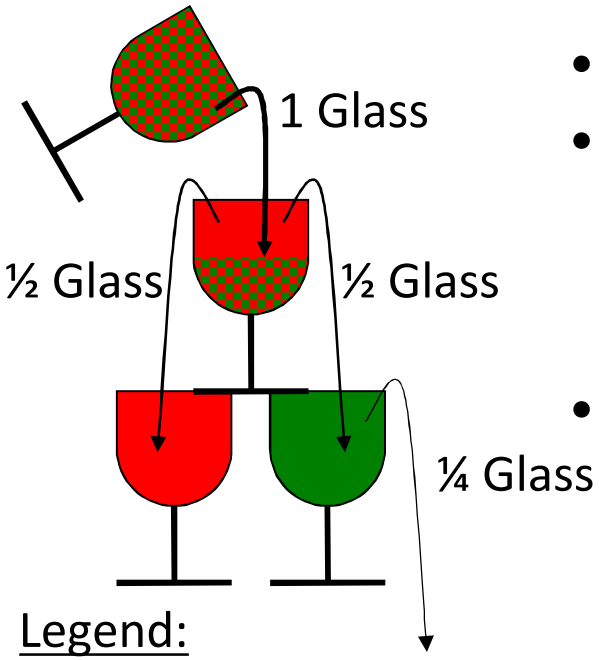


Champagne Sort I (Superstrider Core Algorithm)

Physical model (scalar):

https://cdn.shopify.com/s/files/1/0222/0474/files/champagne-tower-2_grande.jpg?11735

Scalar implementation:



Legend:

Red less than pivot

Green more than pivot

Key Points:

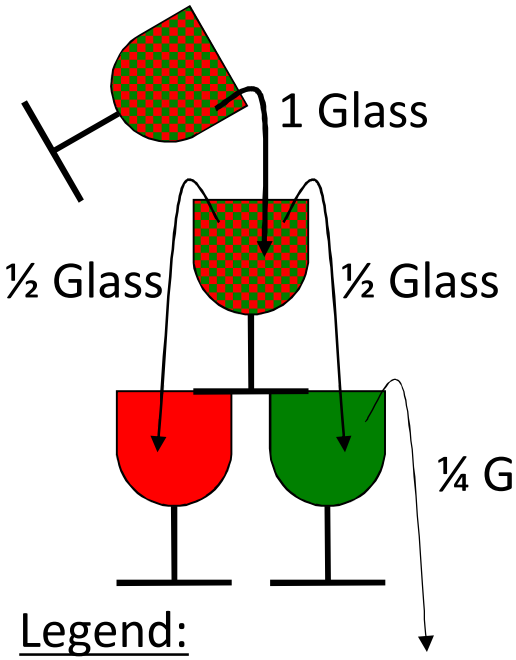
- Glass is DRAM row
- Addition changes just about every DRAM row in the system
- Inefficient

Champagne Sort I (Superstrider Core Algorithm)

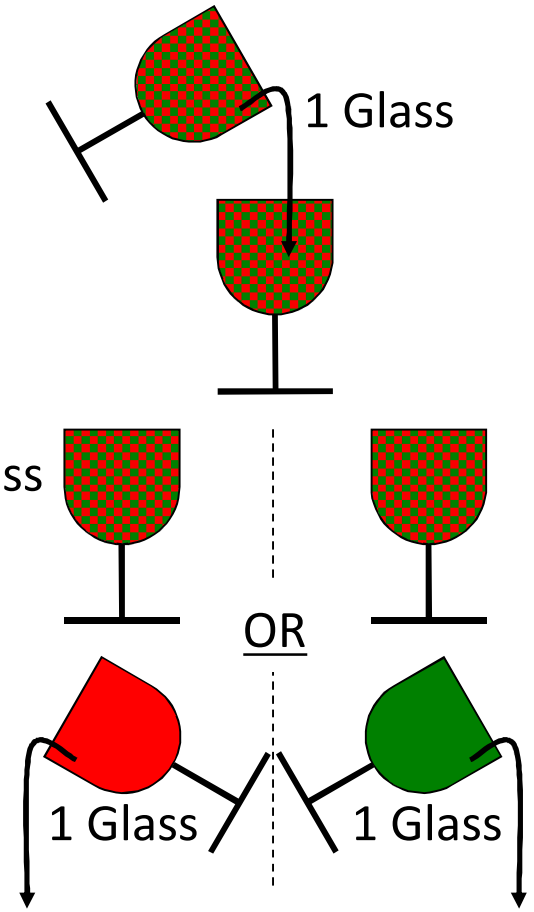
Key points:

- Only have to walk down one side
- Reduces effort a lot

Scalar implementation:



Superstrider (parallel):

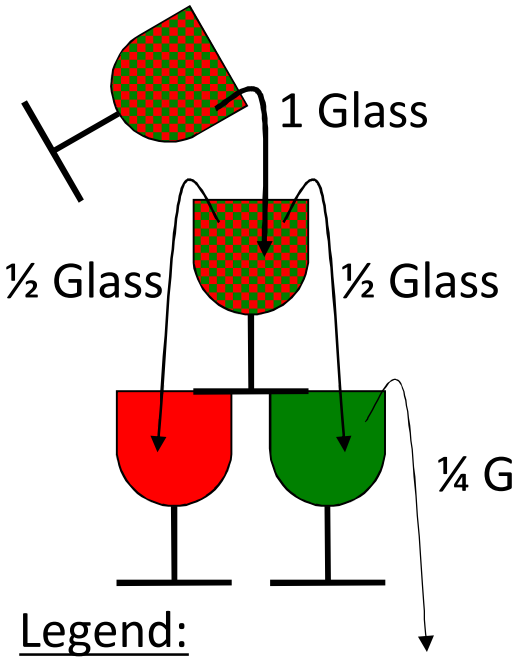


Champagne Sort I (Superstrider Core Algorithm)

Physical model (scalar):

https://cdn.shopify.com/s/files/1/0222/0474/files/champagne-tower-2_grande.jpg?11735

Scalar implementation:

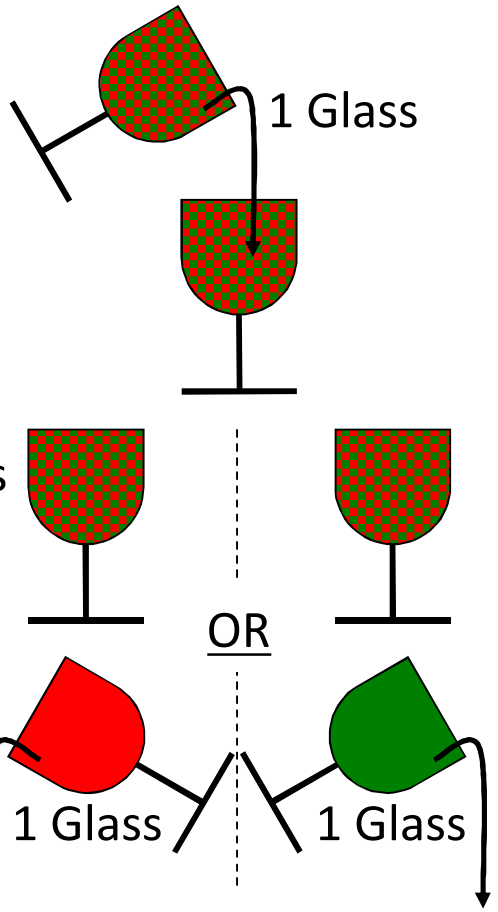


Legend:

Red less than pivot

Green more than pivot

Superstrider (parallel):



Backup: Champagne Sort II (Superstrider Core Algorithm)

- In words
  - Input is a series of records of form
    - { key, value }
  - Superstrider sorts by the key
  - Superstrider also groups of records with the same key and compresses them into a single record
    - { key, Σ values }
  - or a different compression operation than add

Backup: Champagne Sort III (Superstrider Core Algorithm)

Matrix multiply

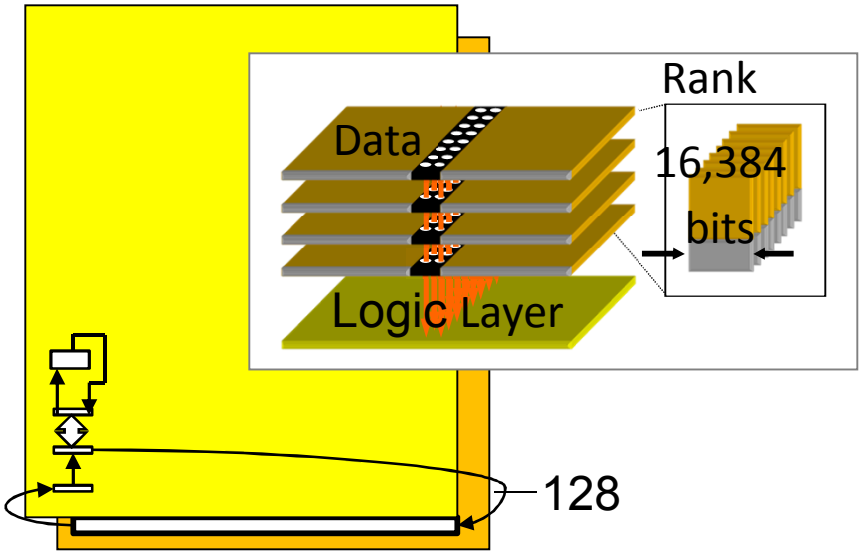
- Definition
  - For matrices *A*, *B*, and *C*
  - $C_{ij} = \Sigma_k A_{ik} * B_{kj}$
- Superstrider
  - Input is $C_{ij}^{(k)} = A_{ik} * B_{kj}$ in random order
  - "Sort" all $C_{ij}^{(k)}$ to the same Champagne class, compressing the value to $C_{ij} = \Sigma_k C_{ij}^{(k)}$

Backpropagation (neural nets)

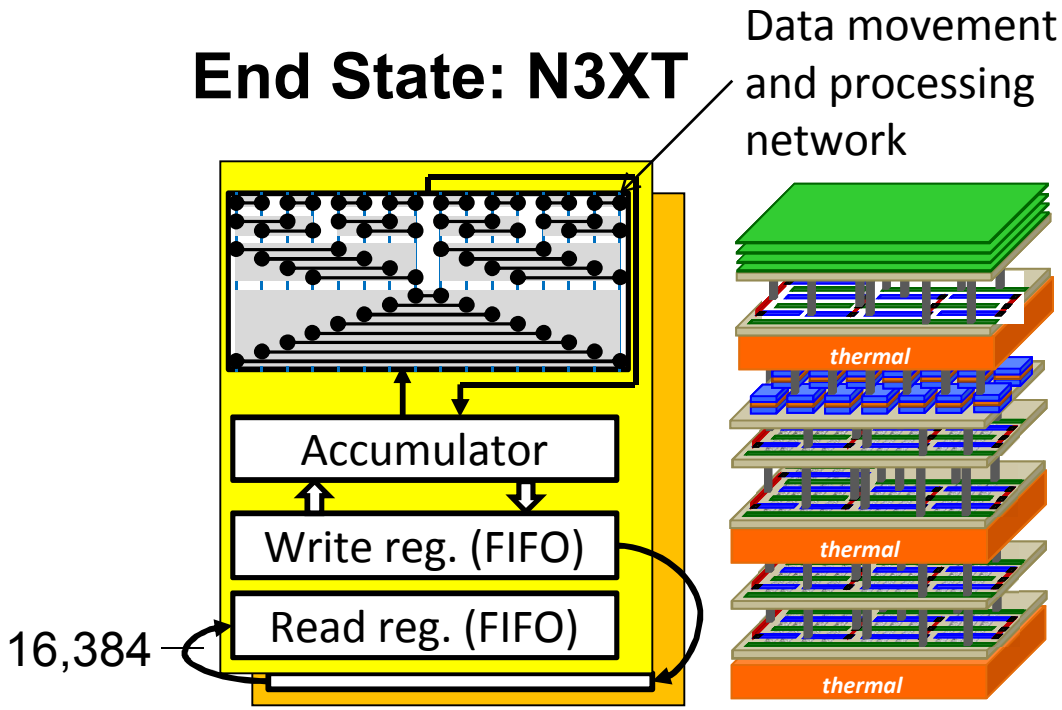
- Definition
  - For weight matrix *W* and vectors *a* and *b*
  - $W'_{ij} = W_{ij} + a_i * b_j$
- Superstrider
  - Create $T_{ji} = \{ W_{ij}, a_i \}$
  - Create $W'_{ij} = f(T_{ji}, b_j)$ where $f(\{w, a\}, b) = w + a*b$

Scale-up with Progressively "Tighter" 3D Integration I

Begin State: HBM (High Bandwidth Memory)

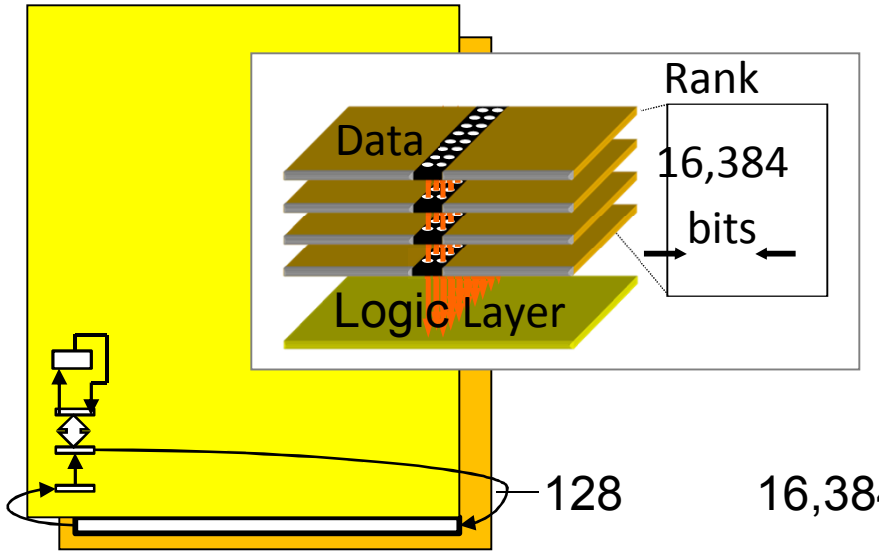


Scale-up with Progressively "Tighter" 3D Integration I

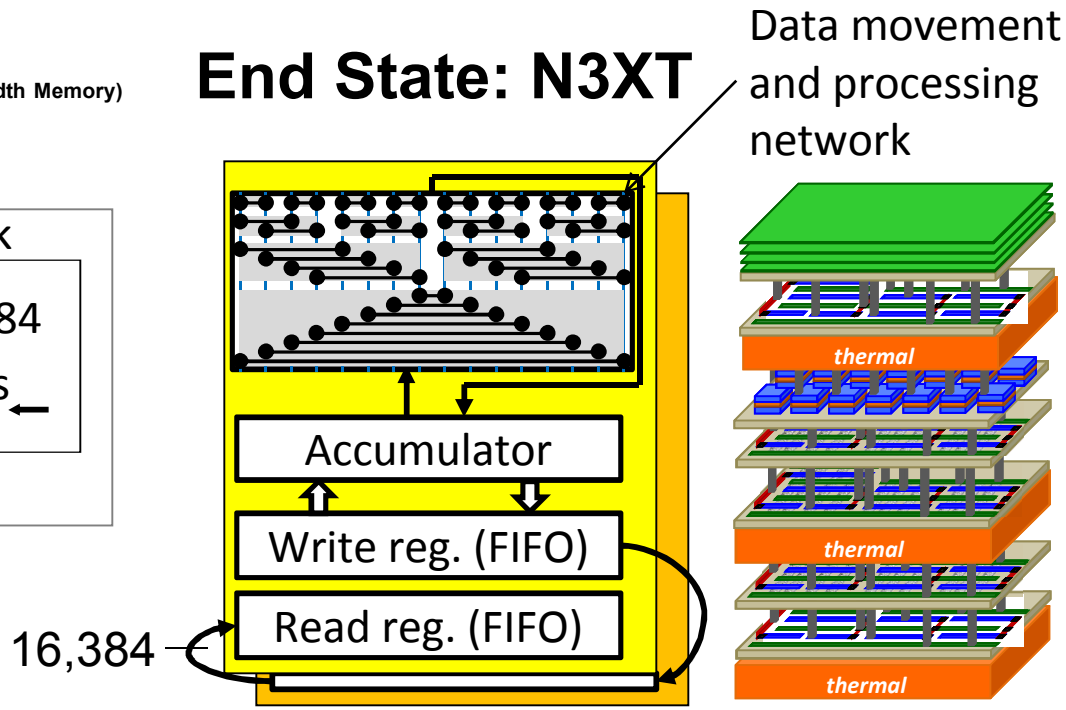


Scale-up with Progressively "Tighter" 3D Integration I

Begin State: HBM (High Bandwidth Memory)



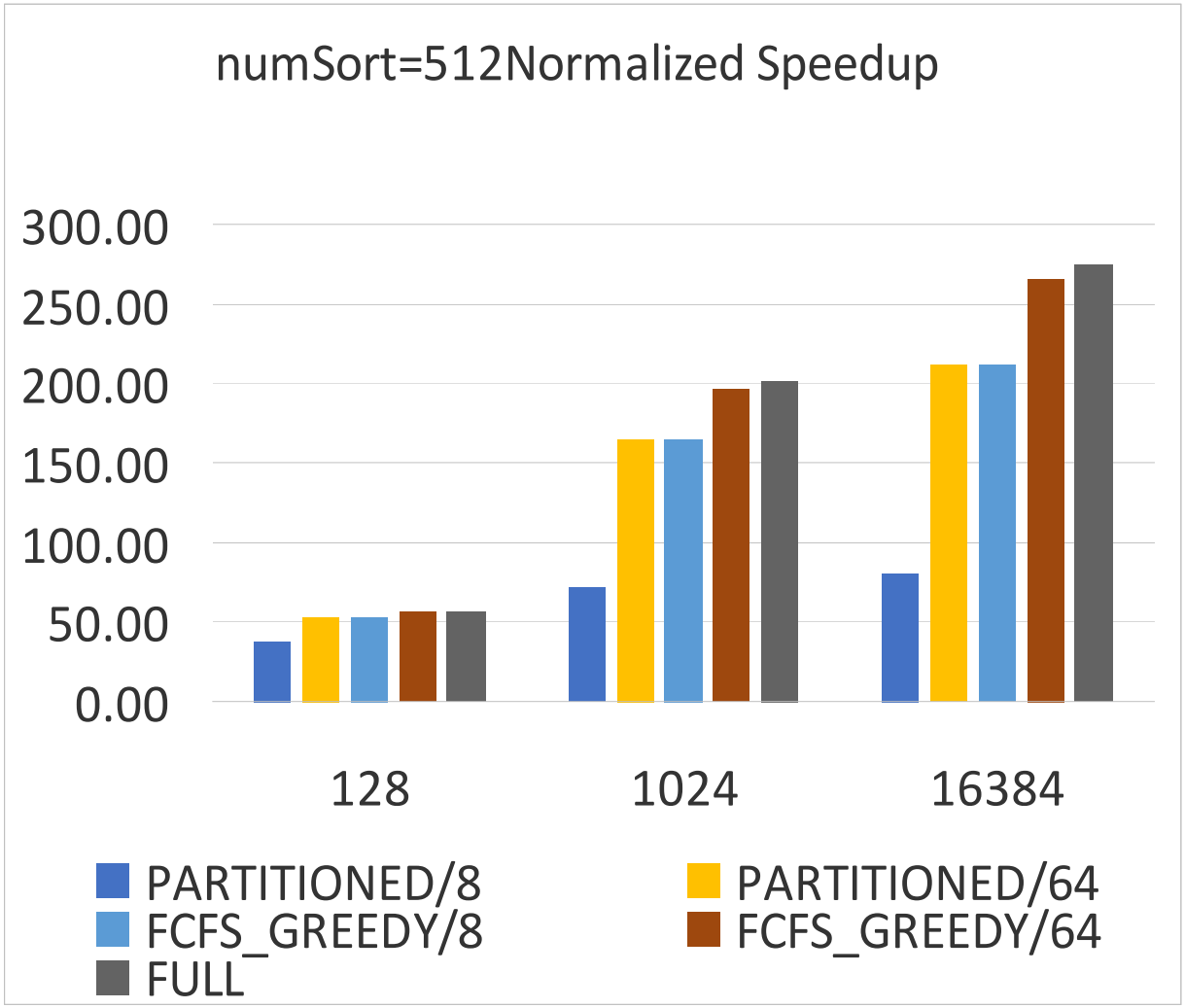
End State: N3XT



HBM Row: 16384 bits =	Bus width	×	Cycles	Network size
(HBM)	128 bits		128	8 = 4 × 2
	256		64	24 = 8 × 3
	512		32	64 = 16 × 4
	1024		16	160 = 32 × 5
(N3XT)	16,384		1	4608 = 512 × 9

Scale-up with Progressively "Tighter" 3D Integration II

- HBM (High Bandwidth Memory)
  - Maybe 50×
  - Immediately actionable
- N3XT (Stanford viewgraph)
  - Maybe 250×
  - Long-term vision
- (Compared to CPU + HBM)



Source: Seshan

Future Directions I

- Accept Top Down Applications Pull
  - Still Moore's law, just shift to Fig. 2 from Fig. 3 in his article

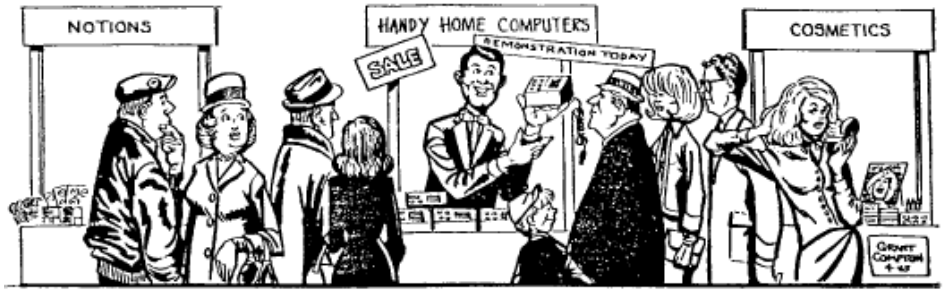


Fig. 2

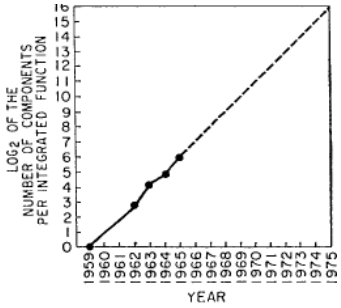


Fig. 3

- Moore, Gordon E. "Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp. 114 ff." *IEEE Solid-State Circuits Society Newsletter* 20.3 (2006): 33-35.

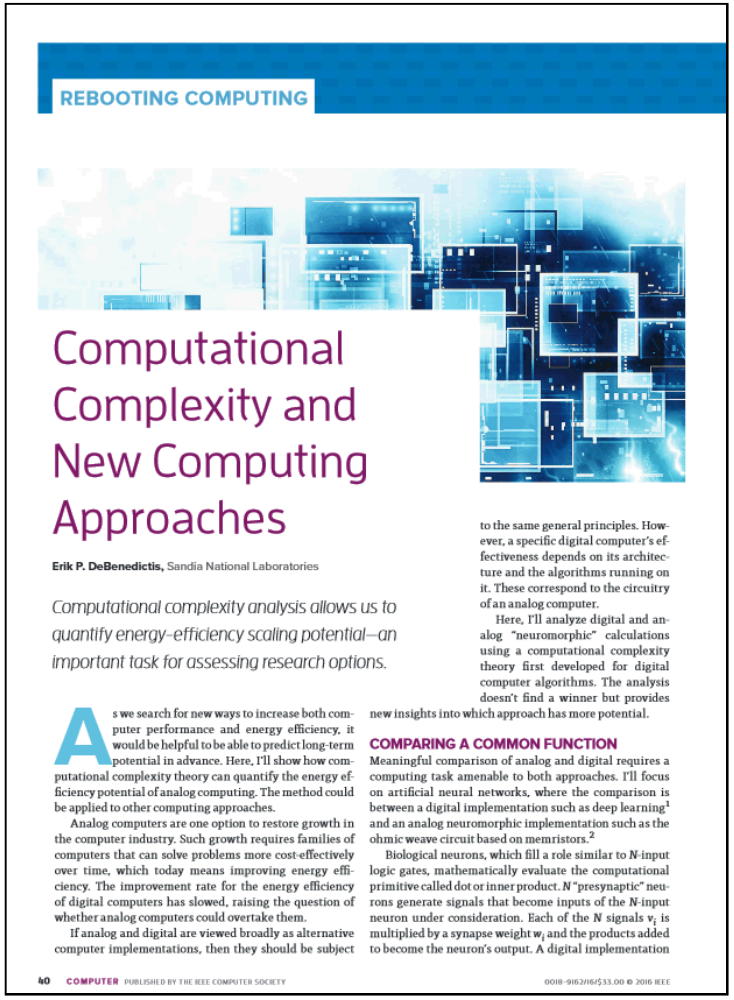
Future Directions II

- IRDS, INC, IEEE Rebooting Computing, and governments should adopt top-down application drive in lieu of bottoms-up technology push
  - Note US Government PCAST report uses term “moonshots” rather than “killer apps”
- Roadmap architectures that can extend scaling
  - in addition to architectures that can enable productive use of new devices
- Superstrider could be developed further
  - Sandia seems interested, but I’m soliciting others

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



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[1] and an analog neuromorphic implementation such as the ohmic weave circuit based on memristors.[2]

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

0018-9162/16/$33.00 © 2016 IEEE

Panel Discussion Questions

- Which technologically enabled innovations will impact society most in the next 10 years?
  - All panelists, 2 foils max
- Which breakthroughs do you anticipate will occur in the next 10 years in AI, computer architecture, devices, technology or any other technical field?
  - Each panelist should select the subjects they like best, 2 foils max
- Which subjects do you think should be addressed by INC in 2018?
  - Each panelist, 2 foils max.
- Which changes (number of days, format, visits, locations etc.) would you like to see in future INC conferences?
  - Each panelist, 2 foils max.

Which tech. innovations will impact society most in the next 10 years?

- AI-class
  - Mobile – introduced by self-driving car but proliferating to other robotics
    - Could be big benefit or create massive unemployment
  - Non-mobile – There is a huge amount of digital data in natural languages (English, Japanese, etc.) that has not been parsed. This could be a big boon for productivity, with an impact on society.
- USG proposes non-commercial “moonshots”
- Other apps?
  - Yeah, but I don’t have an inventory.

PCAST Sample Moonshots

- Bioelectronics for sensory replacement and implantable neuro-stimulation for control of chronic conditions.
- Threat Detection Network
- Distributed Electric Grid
- Global Weather Forecasting

Executive Office of the President, President's Council of Advisors on Science and Technology, Ensuring Long-Term U.S. Leadership in Semiconductors, January 2017, https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_ensuring_long-term_us_leadership_in_semiconductors.pdf

Which breakthroughs in next 10 years?

- AI will make progress
  - Self-driving cars will become safe from existing threats but open new paths for malfeasance exploiting limitation in “IQ” of machines
- Computer architectures will diversify
  - The sequence CPU, FPGA, GPU, KNX, Neural accelerator (e. g. True North) will be extended.
  - Machine learning will start to make architecture an implementation detail
- Devices: 3D proliferates. New memories (e. g. RRAM) will enter production. Better transistors by up to 10×.

INC in 2018?

- If Japanese sponsors like AI/self-driving cars, have conference cover applications pull for advances in hardware, architecture, and software
  - If not AI, there are other areas

Format for INC 2018

- I'm a Rebooting Computing guy
  - How about INC 2018 join the "Rebooting Computing family of advanced computing initiatives"?
- Format from a non-INC guy
  - There is a choice about peer-reviewed papers. INC has not been peer reviewed, but IEEE can help
  - Night before reception, end at noon on second or third day.
  - If Japan is the country, I must profess inadequate knowledge of cities
  - Visits would be a nice option