Serendipity: How Supercomputing Technology is Enabling a Revolution in Artificial Intelligence

José Moreira
IBM Research

HPML/SBAC-PAD 2018 – Lyon, France, September 2018

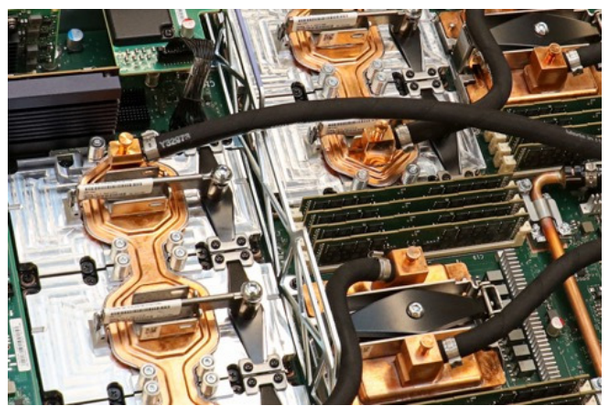
TOP500 – June 2018

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	DOE/SC/Oak Ridge National Laboratory United States	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband IBM	2,282,544	122,300.0	187,659.3	8,806
2	National Supercomputing Center in Wuxi China	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC	10,649,600	93,014.6	125,435.9	15,371
3	DOE/NNSA/LLNL United States	**Sierra** - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband IBM	1,572,480	71,610.0	119,193.6	
4	National Super Computer Center in Guangzhou China	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 NUDT	4,981,760	61,444.5	100,678.7	18,482
5	National Institute of Advanced Industrial Science and Technology (AIST) Japan	**AI Bridging Cloud Infrastructure (ABCI)** - PRIMERGY CX2550 M4, Xeon Gold 6148 20C 2.4GHz, NVIDIA Tesla V100 SXM2, Infiniband EDR Fujitsu	391,680	19,880.0	32,576.6	1,649

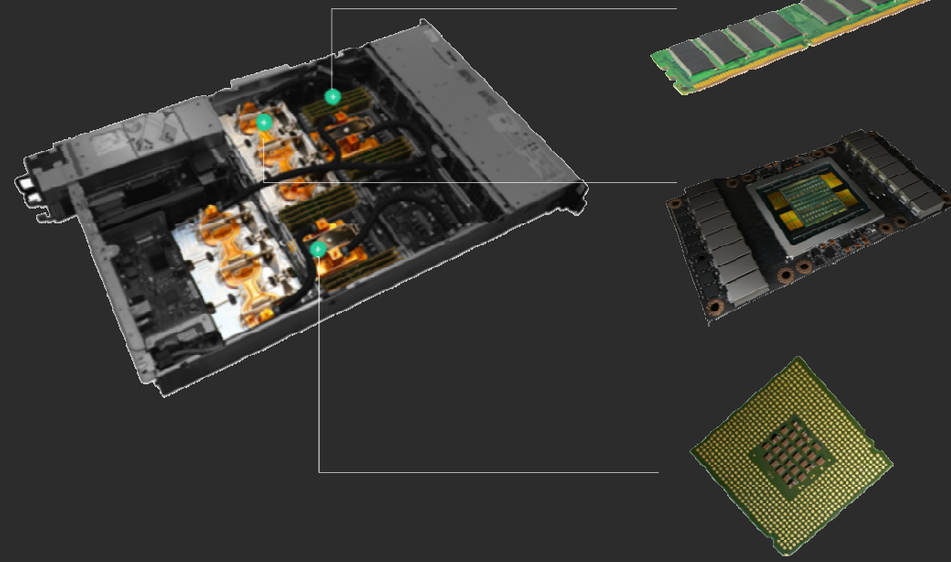
OAK RIDGE National Laboratory | OAK RIDGE LEADERSHIP COMPUTING FACILITY

Summit's Home: DOE ORNL

New 200-Petaflops System is World's Most Powerful and World's Smartest Supercomputer for Science



Next-Gen AI and HPC Platform



256-512 GB DDR4 and 96 GB HBM2

4-6 NVIDIA Volta GPUs

22-core Power9 (2 CPUs per node)

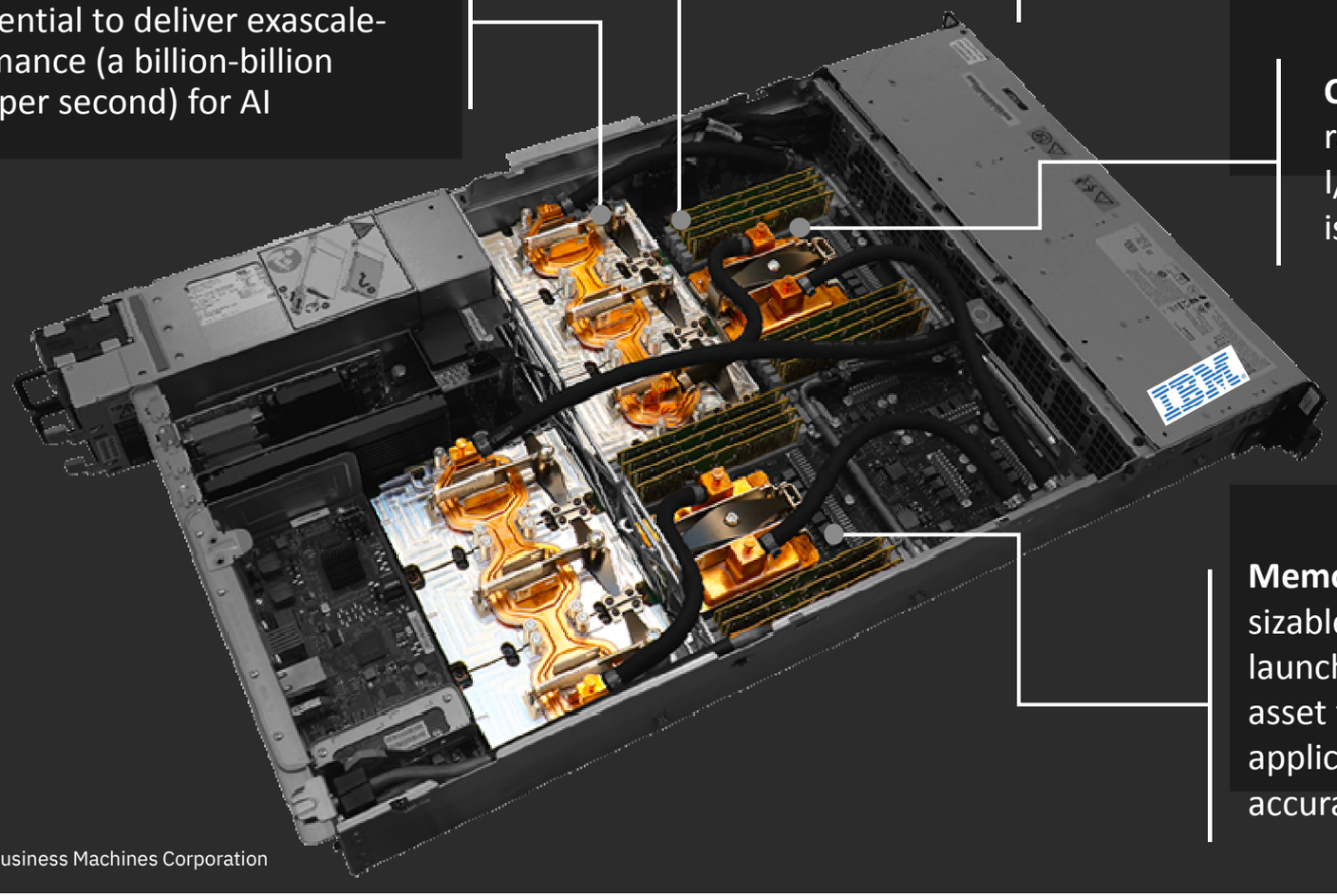
	Summit	Sierra
Peak Performance	**200 Petaflops (DP)**	**125 Petaflops (DP)**
Number of Nodes	4,608	4,320
Node Performance	43/86/700 Teraflops (DP/SP/HP)	29/58/450 Teraflops (DP/SP/HP)
Memory per node	512 GB DDR4, 96 GB HBM2	256 GB DDR4, 64 GB HBM2
NV Memory per node	1600 GB	1600 GB
Total System Memory	**11.1PB** DDR4+HBM2+NV	**9PB** DDR4+HBM2+NV
Compute nodes	**9,216 IBM POWER9™ CPUs** **27,648 NVIDIA Volta™** GPUs	**8,640 IBM POWER9™ CPUs** **17,280 NVIDIA Volta™** GPUs
File System	250 PB, 2.5 TB/s, GPFS™	156 PB, 1.5 TB/s, GPFS™
Power Consumption	**15 MW**	**12 MW**
Interconnect	Mellanox EDR 100G InfiniBand	
Operating System	Red Hat Enterprise Linux (RHEL) version 7.4	

© 2018 International Business Machines Corporation

Key components of Summit/Sierra

GPU Brawn: Summit and Sierra link more than 27,000 and 17,000, respectively, deep-learning optimized NVIDIA GPUs with the potential to deliver exascale-level performance (a billion-billion calculations per second) for AI applications.

High-speed Data Movement: High speed Mellanox interconnect and NVLink high-bandwidth technology built into all of compute processors supply the next-generation information superhighways.



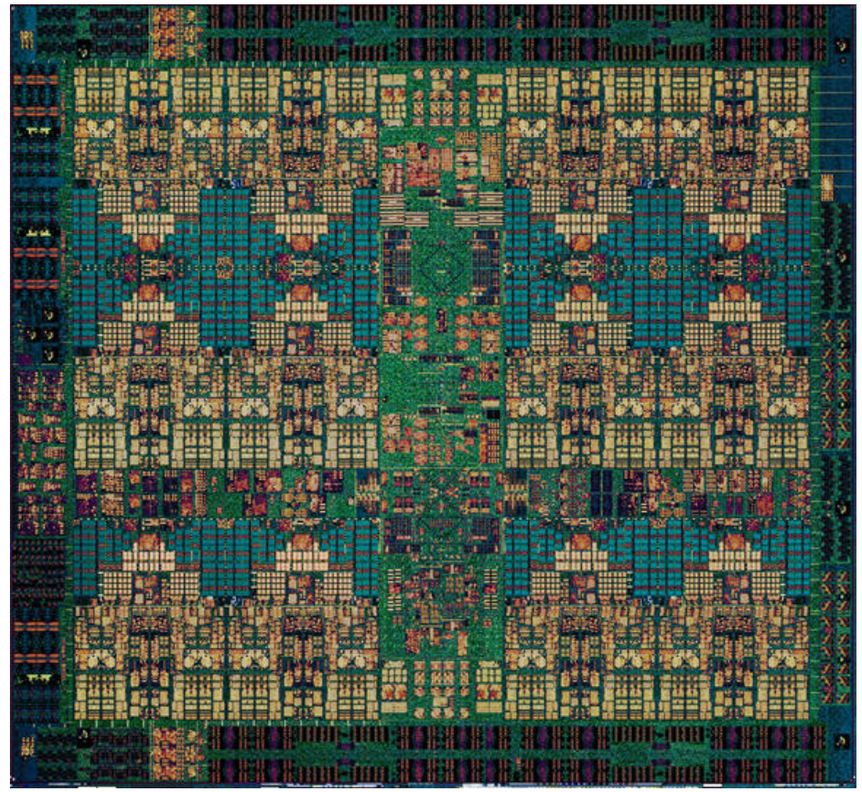
CPU Muscle: IBM Power9 processors to rapidly execute serial code, run storage and I/O services, and manage data so the compute is done in the right place.

Memory Where it Matters: Summit and Sierra's sizable memory gives researchers a convenient launching point for data-intensive tasks, an asset that allows for greatly improved application performance and algorithmic accuracy as well as AI training.

© 2018 International Business Machines Corporation

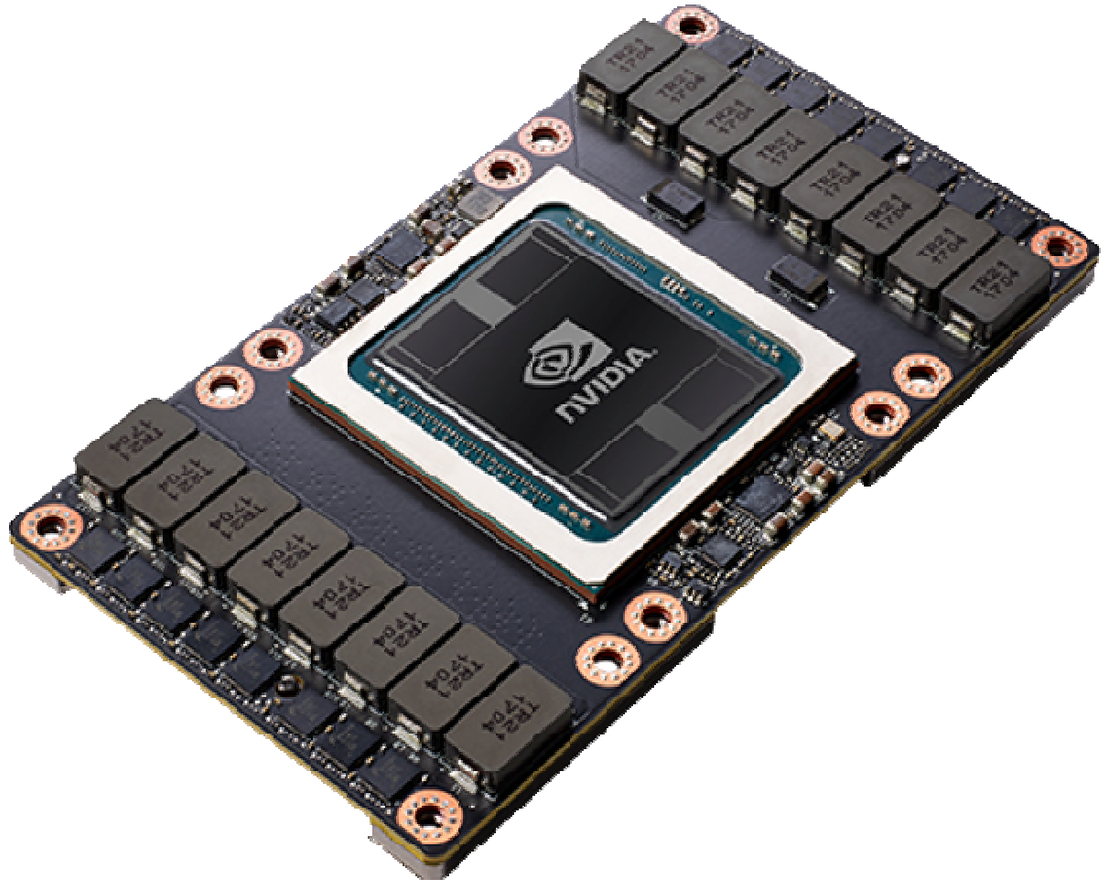
IBM POWER9 Processor

- Up to 24 cores
  - Summit's P9s have 22 cores for yield optimization on first processors
- PCI-Express 4.0
  - Twice as fast as PCIe 3.0
- NVLink 2.0
  - Coherent, high-bandwidth links to GPUs
- 14nm FinFET SOI technology
  - 8 billion transistors
- Cache
  - L1I: 32 KiB per core, 8-way set associative
  - L1D: 32KiB per core, 8-way
  - L2: 256 KiB per core
  - L3: 120 MiB eDRAM, 20-way

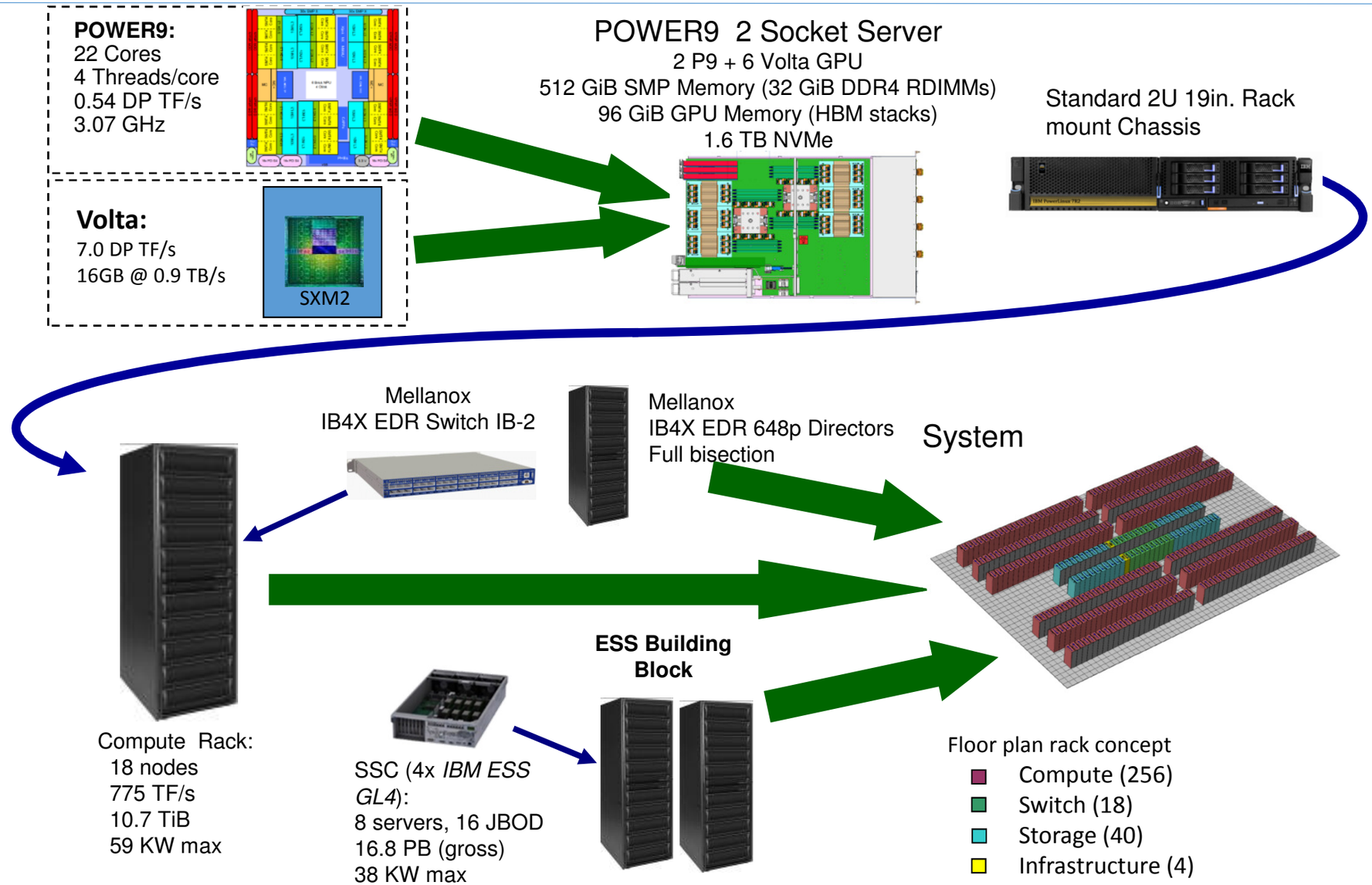


NVIDIA Volta Details

	Tesla V100 for NVLink	Tesla V100 for PCIe
PERFORMANCE with NVIDIA GPU Boost™	DOUBLE-PRECISION 7.8 TeraFLOPS SINGLE-PRECISION 15.7 TeraFLOPS DEEP LEARNING 125 TeraFLOPS	DOUBLE-PRECISION 7 TeraFLOPS SINGLE-PRECISION 14 TeraFLOPS DEEP LEARNING 112 TeraFLOPS
INTERCONNECT BANDWIDTH Bi-Directional	NVLINK 300 GB/s	PCIE 32 GB/s
MEMORY CoWoS Stacked HBM2	CAPACITY 16 GB HBM2 BANDWIDTH 900 GB/s	



Summit 200 PetaFlops System – June, 2018



5/6 Gordon Bell Finalists at SC18 (Summit/Sierra)

1. Uncovering hidden networks of genes by comparing genetic variations within a population (ORNL) – 2.36 exaops!
2. Applying artificial intelligence and mixed-precision arithmetic to the simulation of earthquake physics in urban environments (University of Tokyo)
3. Using deep neural networks to identify extreme weather patterns from high-resolution climate simulations (LBNL) – 1.13 exaops!
4. Intelligent software that can automatically identify materials' atomic-level information from electron microscopy data (ORNL) – 152.5 petaflops
5. Improved lattice quantum chromodynamics algorithms to help scientists predict the lifetime of neutrons and answer fundamental questions about the universe (LBNL/LLNL) - 15× over Titan

https://www.olcf.ornl.gov/2018/09/17/uncharted-territory/

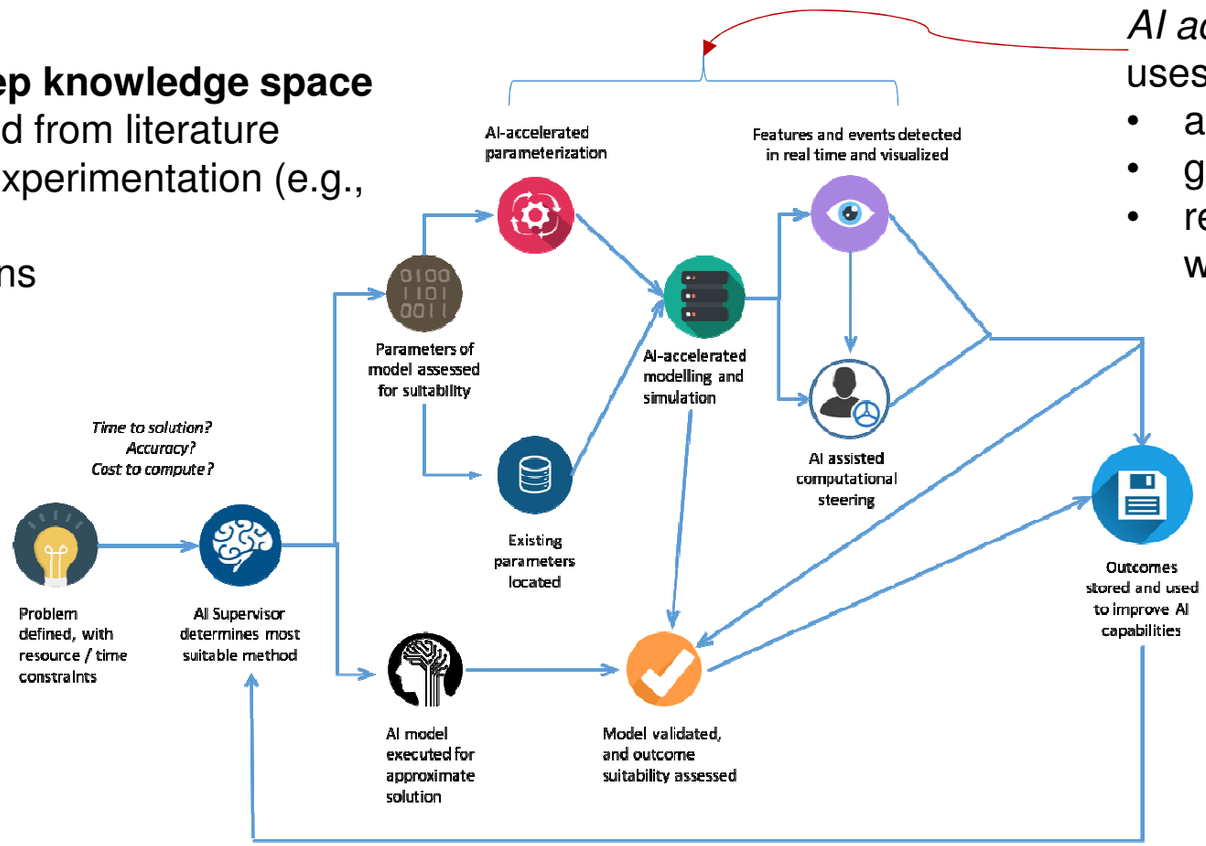
Overview: Cognitive discovery / Intelligent simulation

Cognitive Discovery depends on the ingestion and representation of knowledge, be that from the literature, experiment, or from previously run simulations to develop a **deep knowledge space**

Intelligent Simulation uses **machine learning techniques** to optimize ensemble analysis, speed convergence, improve fidelity, increase numerical stability, and actively steer simulations to speed up time to solution.

AI supervisor uses **deep knowledge space**

- knowledge extracted from literature
- data derived from experimentation (e.g., wet labs)
- data from simulations



AI accelerated modelling and simulation uses **machine learning techniques** to

- accelerate parameterization of models
- guide ensembles of simulations
- replace expensive simulation routines with AI-derived approximations

The Unreasonable Effectiveness of Mathematics in the Natural Sciences

- *Eugene Wigner*, 1960:

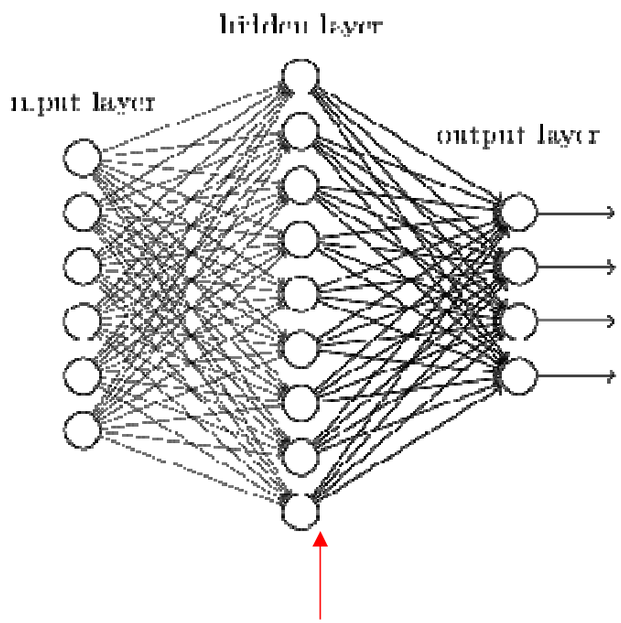
  "The miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics is a wonderful gift which we neither understand nor deserve. We should be grateful for it and hope that it will remain valid in future research and that it will extend, for better or for worse, to our pleasure, even though perhaps also to our bafflement, to wide branches of learning."

- How about the effectiveness of HPC hardware for ML/AI?

  That one is a little easier to explain … ☺

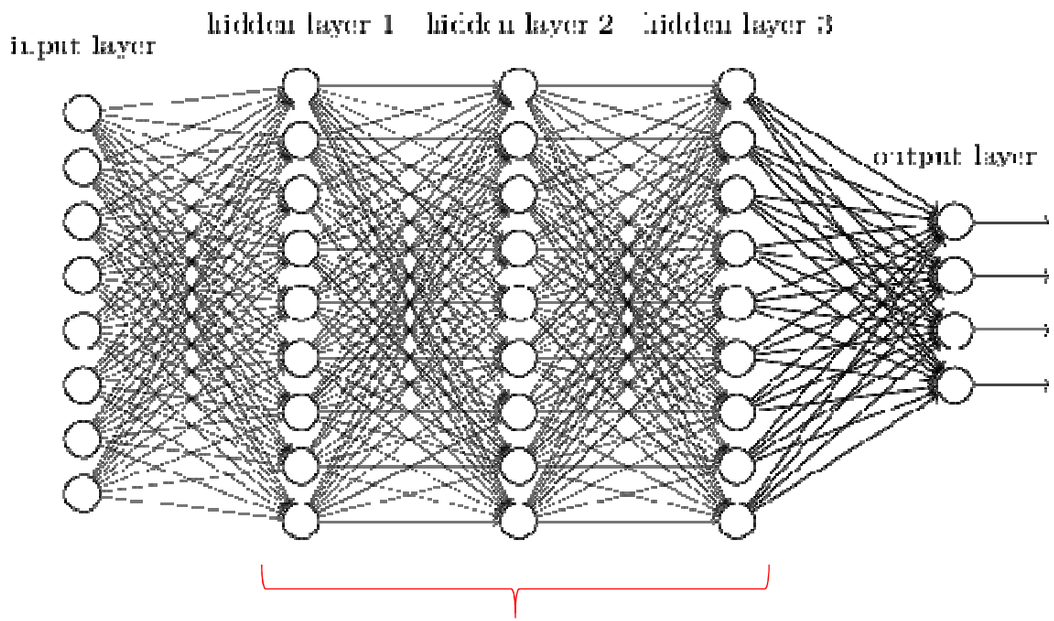
The serendipity between GPUs and DNNs

Shallow neural network



1 hidden layer

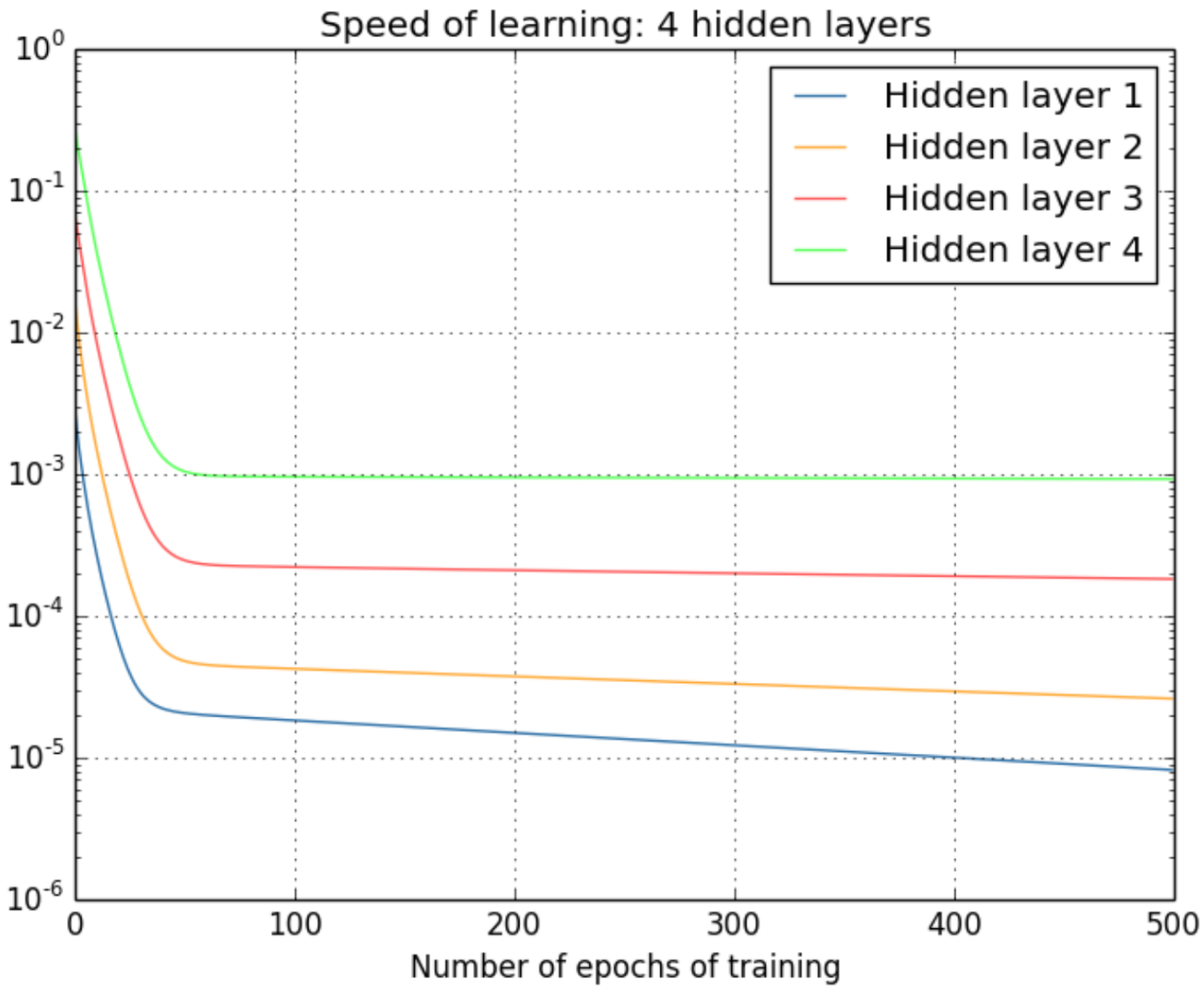
Deep neural network



2+ hidden layers

Any questions?

Deep neural networks are hard to train



Gradient instabilities!

Hard ≠ impossible

- As described in a 2010 paper, Cireșan, Meier, Gambardella, and Schmidhuber* have trained a flat, 5-hidden layer DNN (2500, 2000, 1500, 1000, and 500 neurons, respectively – approximately 12 million weights) to classify the MNIST dataset
- This required no techniques beyond what was known in the 1980's, but consumed a lot of compute power, achieved through GPUs
- They achieved a classification accuracy of 99.65%! (that is, 35 misclassified images)

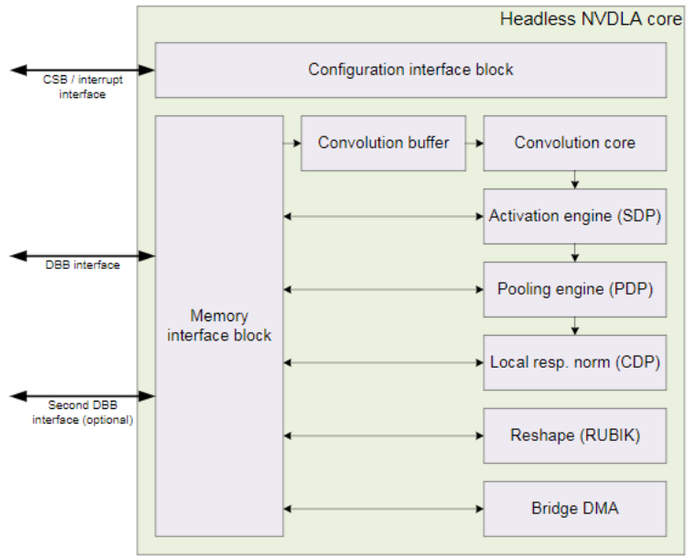
*Deep, Big, Simple Neural Nets Excel on Handwritten Digit Recognition, by Dan Claudiu Cireșan, Ueli Meier, Luca Maria Gambardella, and Jürgen Schmidhuber (2010).

Divergence brewing?

Tesla V100 GPU

	Tesla V100 for NVLink	Tesla V100 for PCIe
PERFORMANCE with NVIDIA GPU Boost™	DOUBLE-PRECISION 7.8 TeraFLOPS	DOUBLE-PRECISION 7 TeraFLOPS
	SINGLE-PRECISION 15.7 TeraFLOPS	SINGLE-PRECISION 14 TeraFLOPS
	DEEP LEARNING 125 TeraFLOPS	DEEP LEARNING 112 TeraFLOPS
INTERCONNECT BANDWIDTH Bi-Directional	NVLINK 300 GB/s	PCIE 32 GB/s
MEMORY CoWoS Stacked HBM2	CAPACITY 16 GB HBM2	
	BANDWIDTH 900 GB/s	

NVIDIA Deep Learning Accelerator

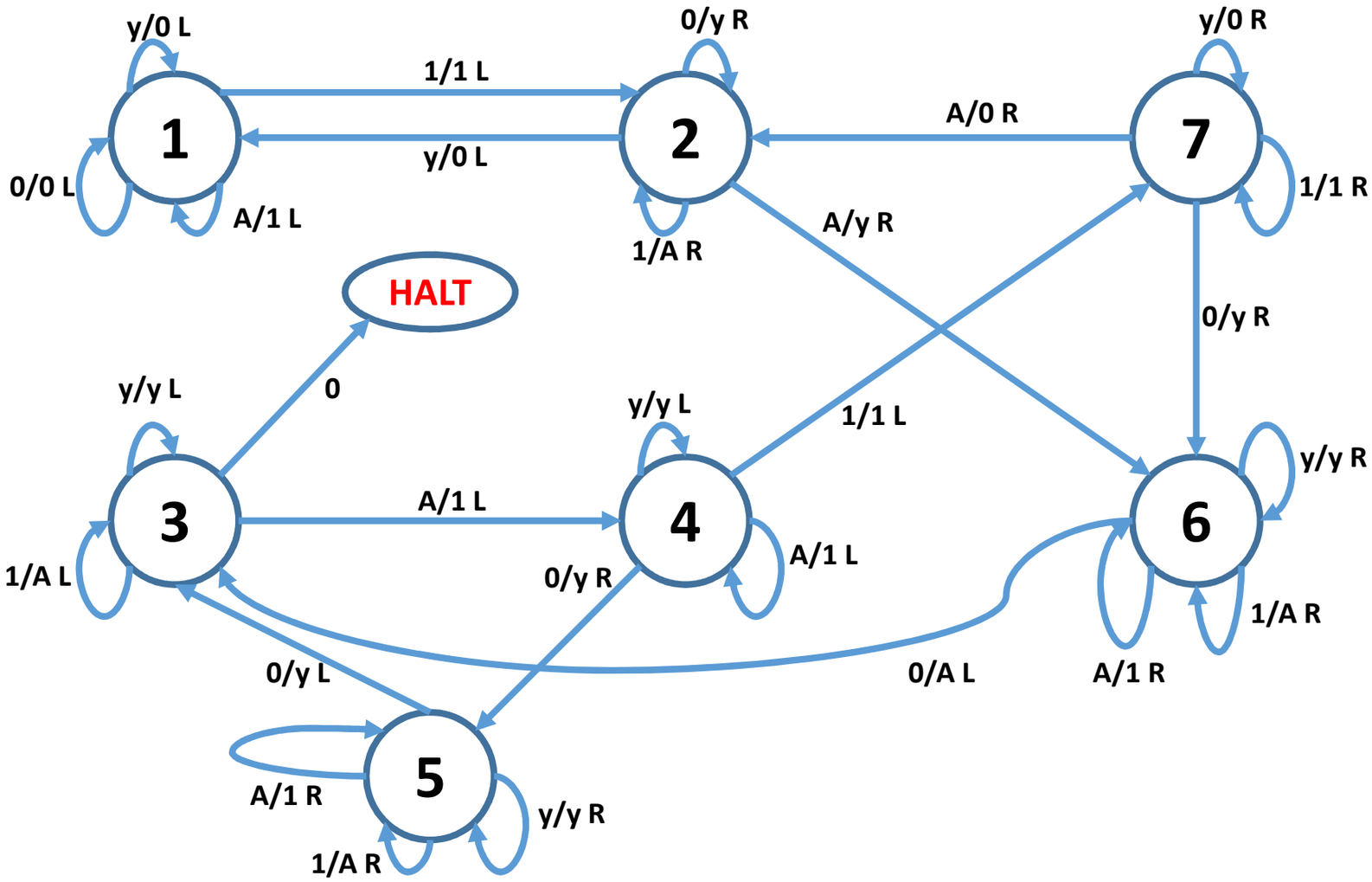


Tesla T4 GPU (int8/int4)

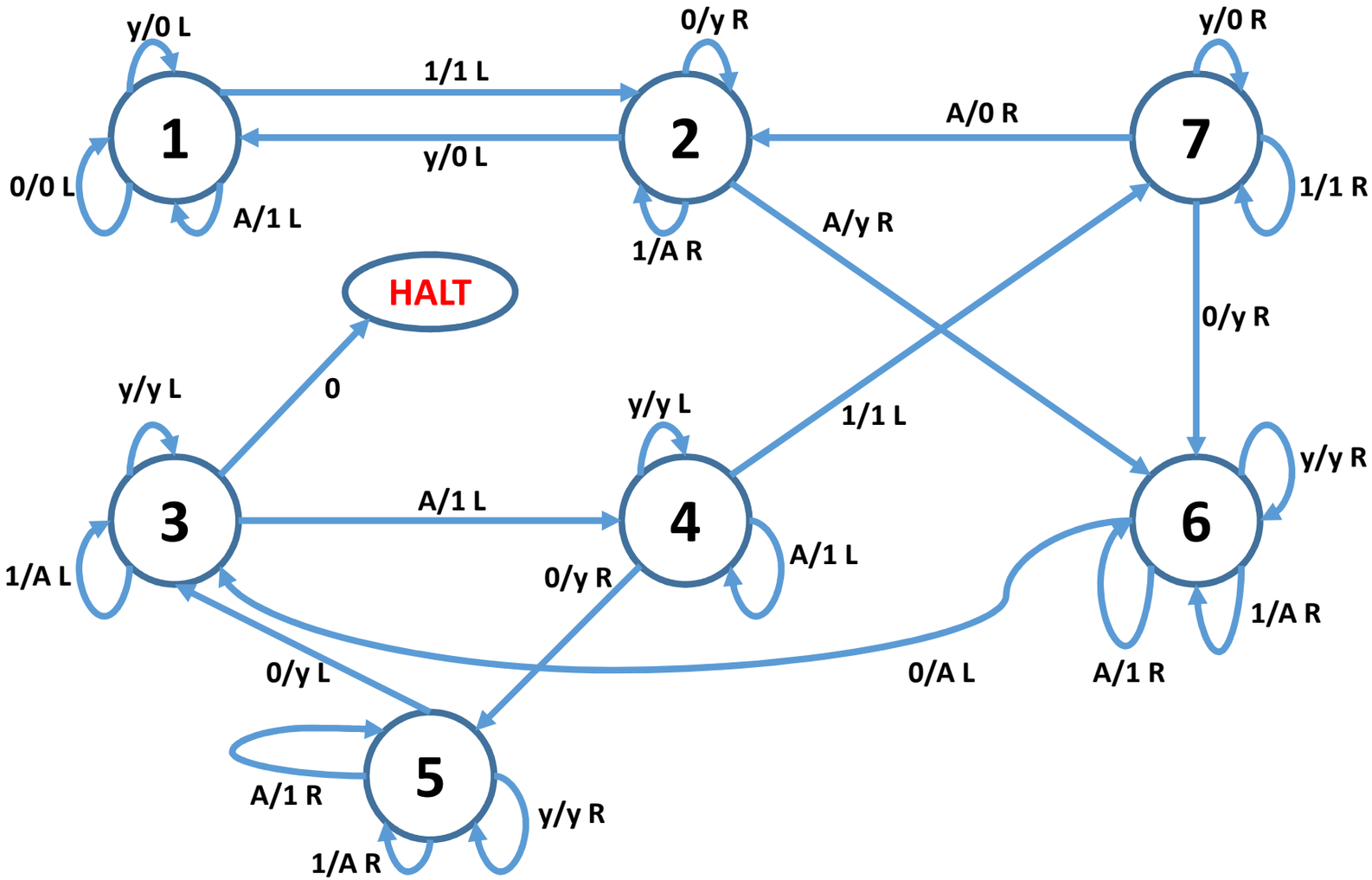
Machine consciousness

- Our machines are approaching the (suspected) processing power of the human brain
- We have been able to train and deploy deep neural networks of a scale unimaginable just a decade ago
- We continue to make progress at better algorithms and better implementation of large machine learning solutions
- Does that mean we are close to achieving *General Artificial Intelligence*?
- If so, when do we declare such machines thinking/conscious?

Quiz: What is this?

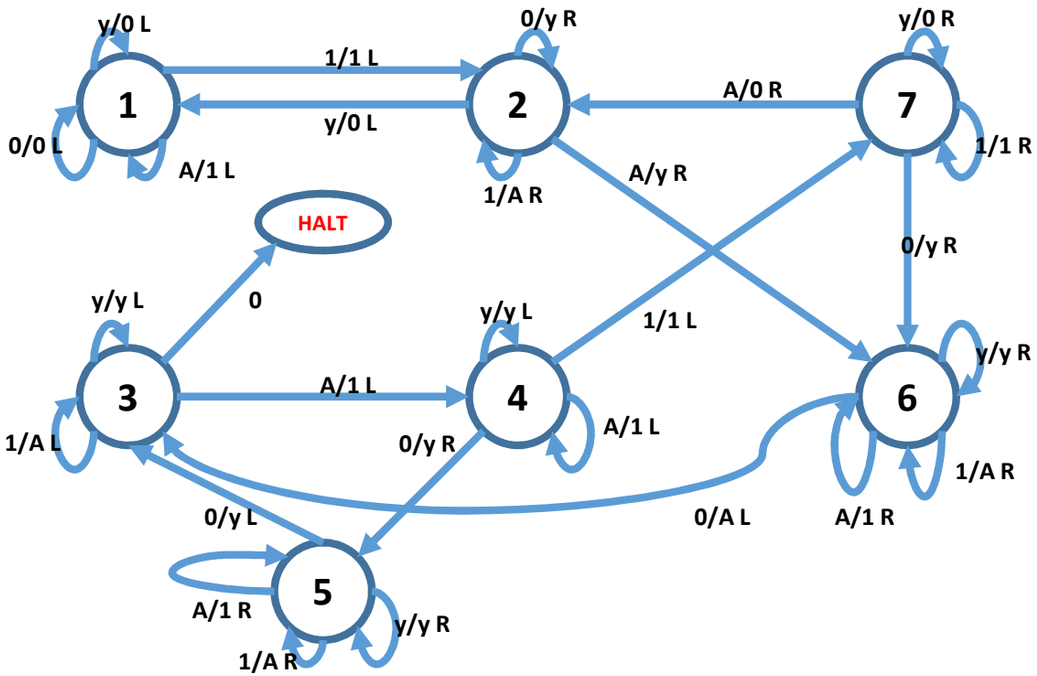


Quiz: What is this? (Marvin Minsky)

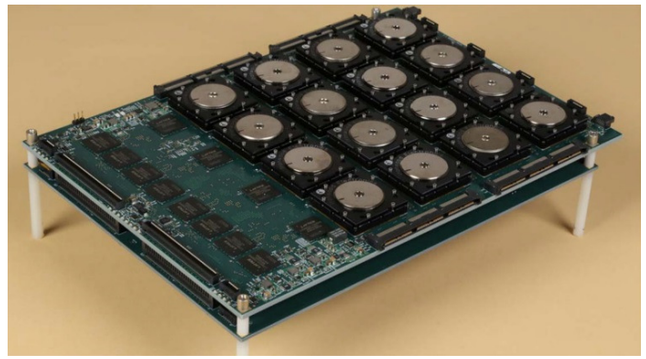


Answer: A (Minsky's) *universal Turing machine (minus the tape)*

Equivalent to any machine we know how to build



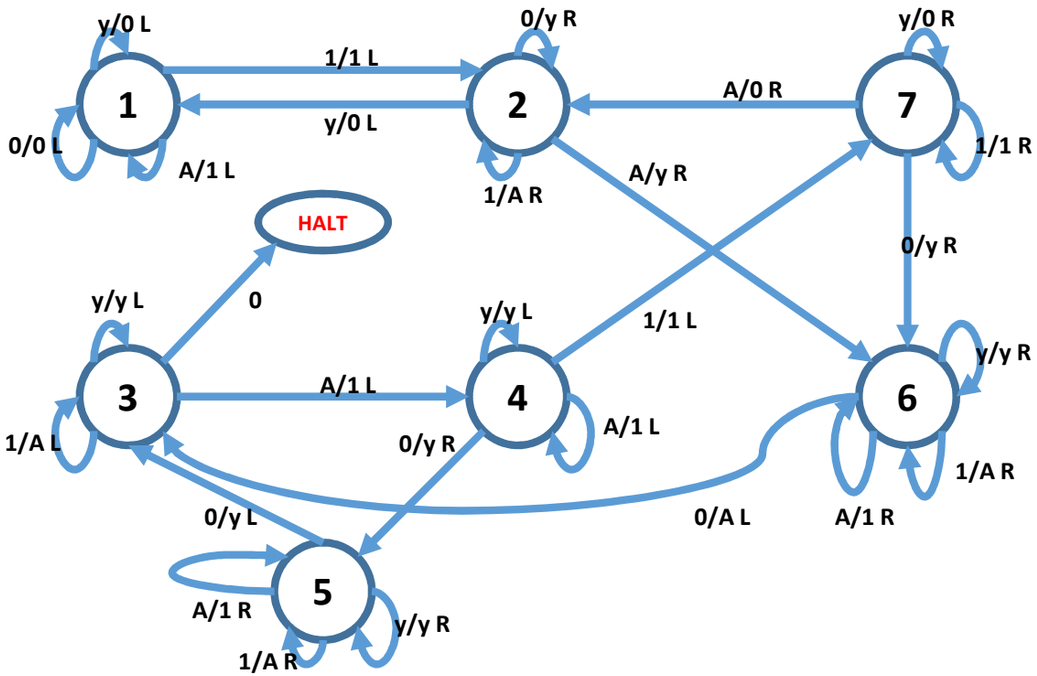
≡



Not equivalent to everything our minds can conceive

- Let $T = \{T_i : T_i \text{ is a Turing machine}\}$
- Let $I = \{I_j : I_j \text{ is an input sequence to a Turing machine}\}$
- $|T| = \aleph_0$
- $|I| = \aleph_0$
- Let $C = \{C_i(j) : C_i(j) \text{ is the computation perfomed by } T_i \text{ on } I_j\}$
- $|C| = \aleph_0$
- $|\mathbb{R}| = \aleph_1 = 2^{\aleph_0}$, that is, almost all real numbers are noncomputable

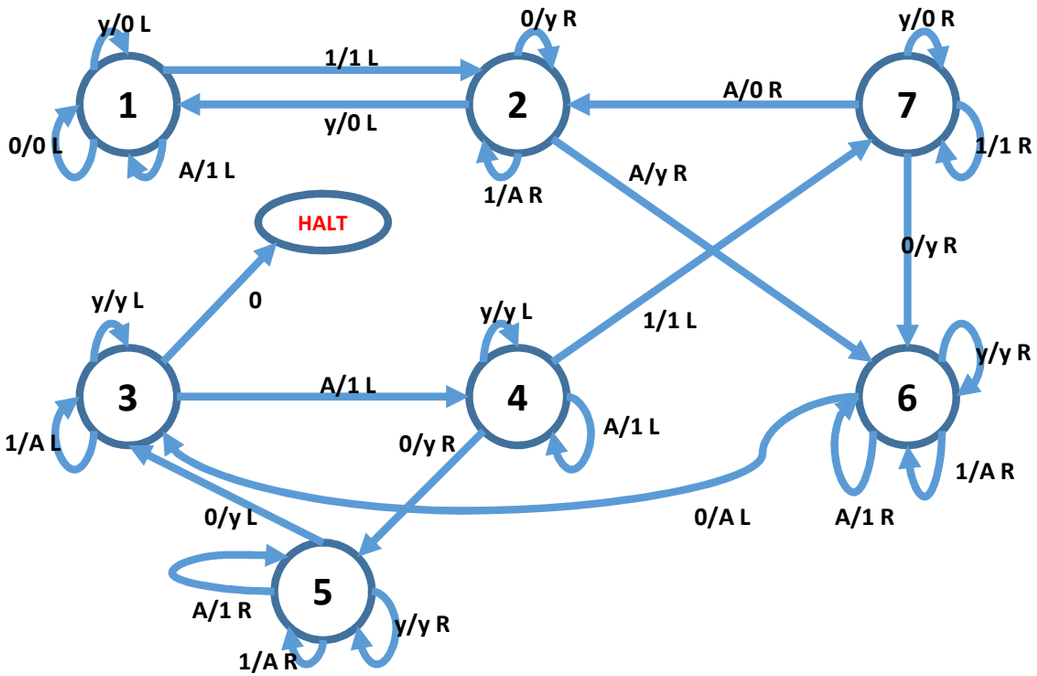
But, is it equivalent to everything that exists?



≡

?

But, is it equivalent to everything that exists?



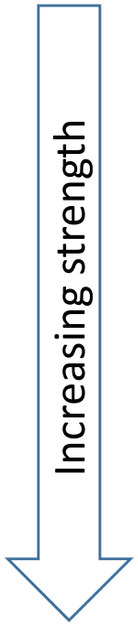
≡

?

Answer: We don't know!

Hypothesis of Digital Physics

1. The physical world is essentially informational
2. The physical world is essentially computable
3. The physical world can be described digitally
4. The physical world is in essence digital
5. The physical world is itself a computer
6. The physical world is the output of a simulated reality exercise

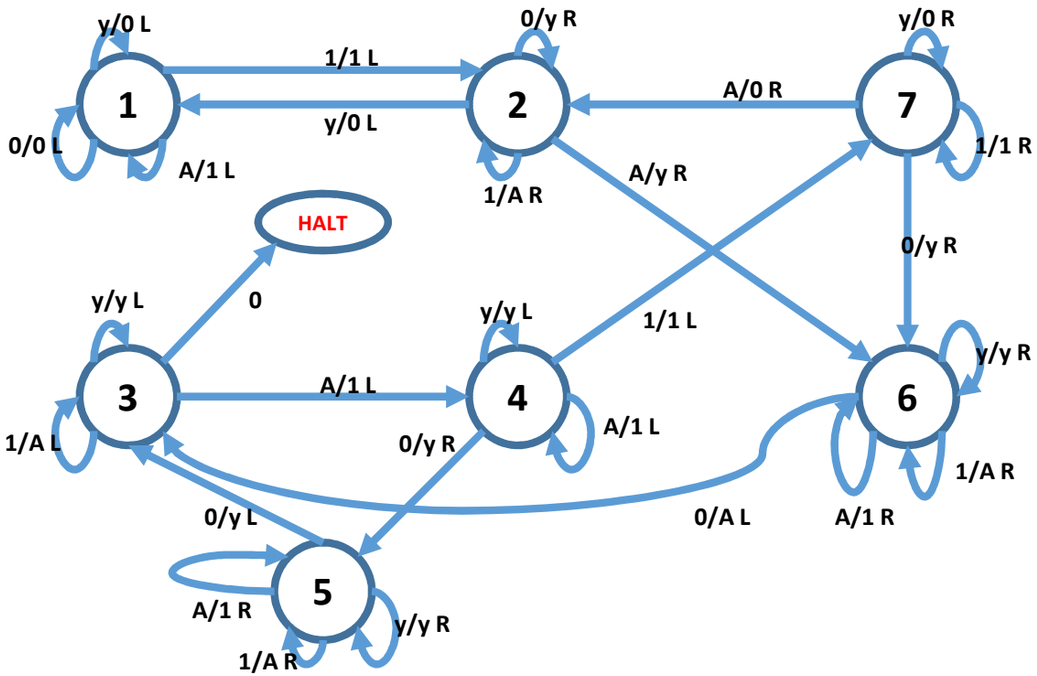


Computable universe hypothesis (CUH) – Max Tegmark

A Bunch of Rocks

https://xkcd.com/505/

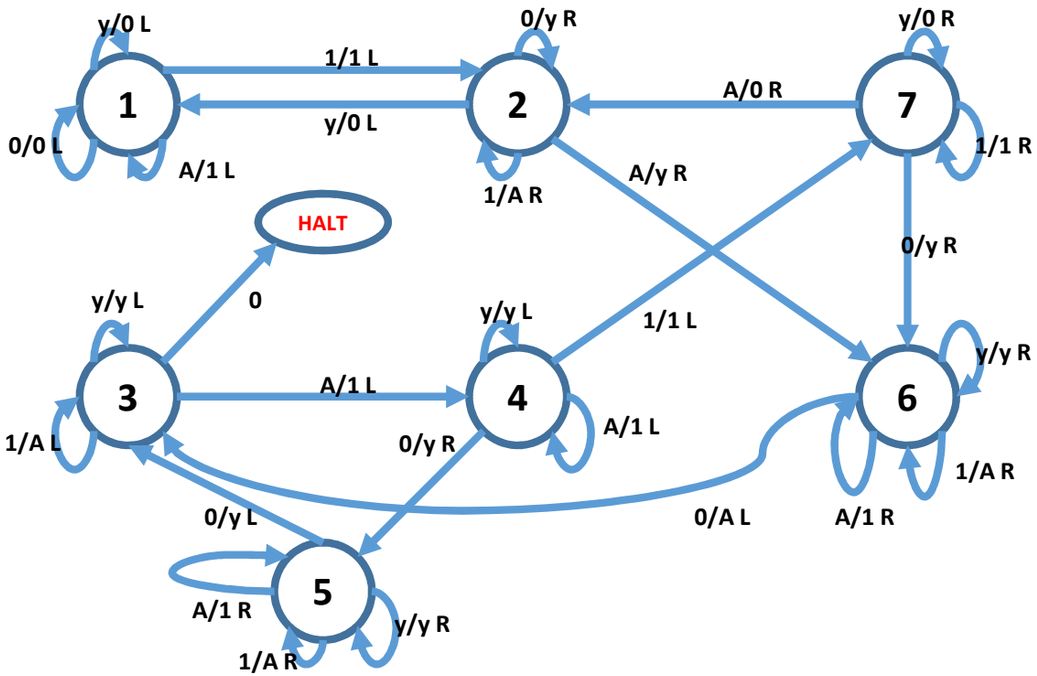
More relevant to cognitive computing …



≡

?

More relevant to cognitive computing …



≡

?

Answer: We don't know!

Viewpoints of consciousness (Roger Penrose)

A. All thinking (including consciousness) is computation – *strong* or *hard* Artificial Intelligence
B. Consciousness is a physical process that can be simulated computationally (known physics), but simulation cannot by itself evoke consciousness (simulation of a process ≠ process itself) – *weak* or *soft* Artificial Intelligence

Ex Machina (2015 film – A *vs* B)

Viewpoints of consciousness (Roger Penrose)

A. All thinking (including consciousness) is computation – *strong* or *hard* Artificial Intelligence
B. Consciousness is a physical process that can be simulated computationally (known physics), but simulation cannot by itself evoke consciousness (simulation of a process ≠ process itself) – *weak* or *soft* Artificial Intelligence
C. Consciousness is a physical process that cannot be simulated computationally – new physics
D. Consciousness cannot be scientifically explained – mysticism

Proof (?) that there is more than computation (RP)

- Let A be an algorithmic procedure taking two inputs, i and j, such that if $A(i,j)$ terminates then $C_i(j)$ does not terminate
- A is known and believed to be sound – encapsulates all the procedures available for demonstrating that computations do not stop
- It does not need to be perfect – $A(i,j)$ may not terminate when $C_i(j)$ does not terminate
- Make $i = j$, then if $A(j,j)$ terminates then $C_j(j)$ does not terminate
- But $A(j,j) = C_k(j)$, for some k
- Make $j = k$, then if $A(k,k)$ terminates then $C_k(k)$ does not terminate
- But $A(k,k) = C_k(k)$
- Then if $A(k,k)$ terminates then $A(k,k)$ does not terminate
- Therefore, $A(k,k)$ does not terminate – does not encompass all our knowledge!

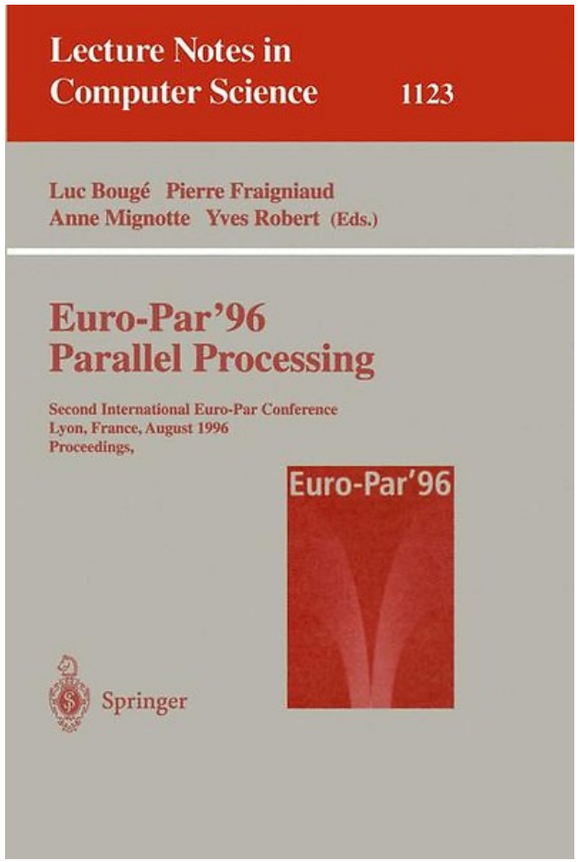
Conclusions

- The new Summit/Sierra supercomputers are ushering a new era of large-scale scientific/artificial intelligence computing
- Even more significant, scientific/artificial intelligence computing are being integrated into new large-scale applications
- Will integration at the application level continue to drive integration at the hardware level, or are we going to see more divergence?
- Current machine learning/cognitive computing research will not give us thinking/conscious machines
- And it doesn't have to – that is not the purpose
- "Is the human mind computable?" – ~~*One of*~~ *the most important scientific questions we face today!*

Bibliography/Further reading

- Marvin Minsky. **Computation: Finite and Infinite Machines**. *Prentice-Hall*.
- Roger Penrose. **The Large, the Small and the Human Mind**. *Cambridge University Press.*
- Roger Penrose. **Shadows of the Mind**. *Oxford University Press*.
- Hava T. Siegelmann. **Computation Beyond the Turing Limit**. *Science*, vol. 268, no. 5210.
- Robert I. Soare. **Turing Oracle Machines, Online Computing, and Three Displacements in Computability Theory**. http://www.people.cs.uchicago.edu/~soare/History/turing.pdf.
- Federico Faggin. **What is Consciousness?** http://www.fagginfoundation.org/articles-2/.
- Max Tegmark. **Our Mathematical Universe: My Quest for the Ultimate Nature of Reality**. *Penguin Random House*.
- Masafumi Oizumi, Larissa Albantakis, Giulio Tononi. **From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0**. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003588

On a personal note …



22 years!