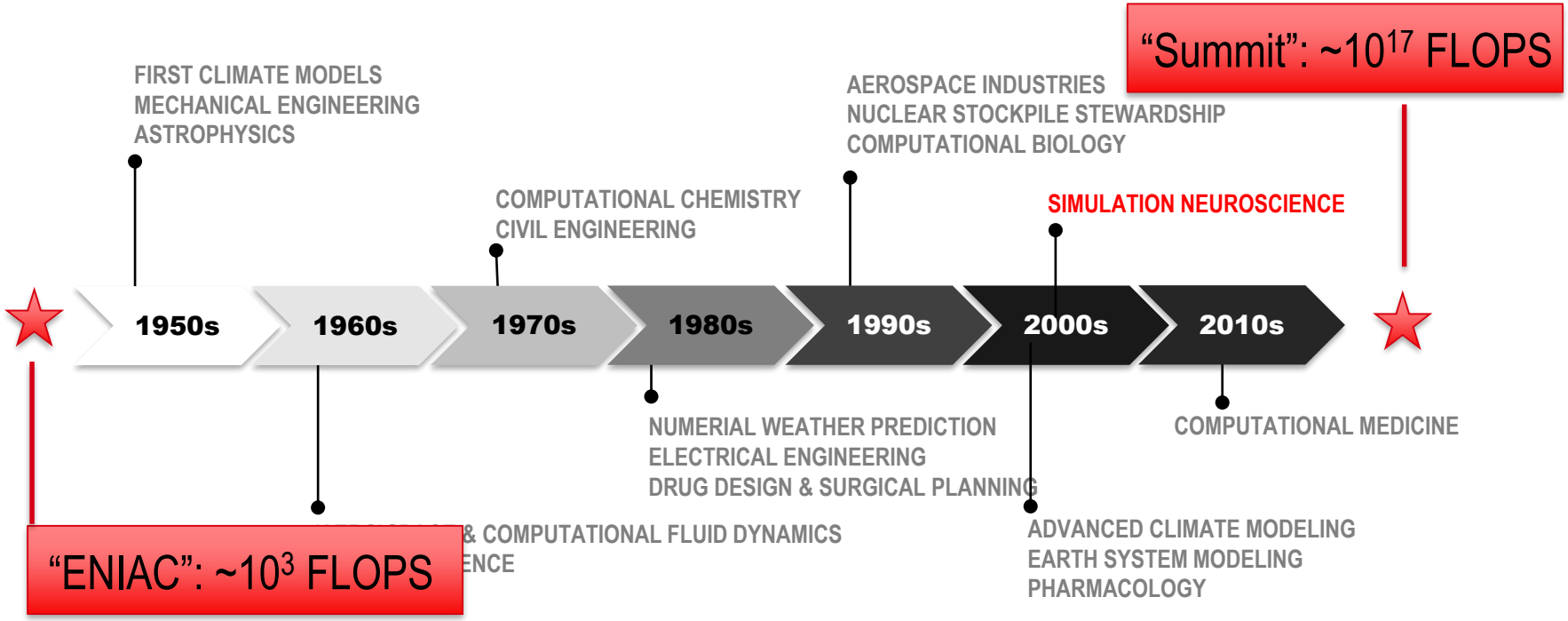
Neuroscience and the Future of Computing

Prof. Felix Schürmann
Blue Brain Project – Co-Director, Head of the Computing Division
Adjunct Professor, Laboratory for Neurosimulation Technology
Ecole polytechnique fédérale de Lausanne, CH

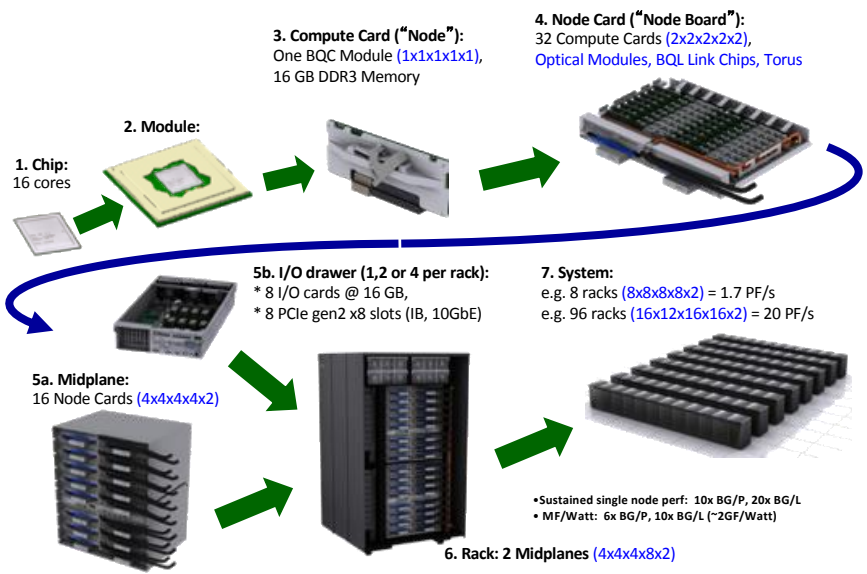
Simulation Science



ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Blue Brain Project

Let's Remind Ourselves – what is a supercomputer?



Source: IBM

OLCF – "Summit"

Sponsors	U.S. Department of Energy
Operators	IBM
Architecture	9,216 POWER9 22-core CPUs 27,648 Nvidia Tesla V100 GPUs[1]
Power	13 MW[2]
Storage	250 PB
Speed	200 petaflops (peak)
Purpose	Scientific research
Web site	www.olcf.ornl.gov/olcf-resources/compute-systems/summit/

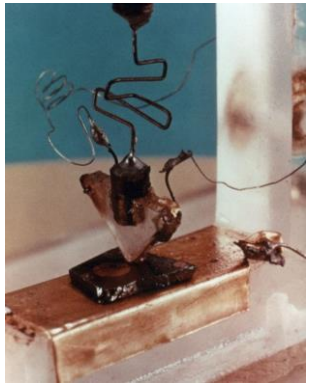
2,397,824 cores!!

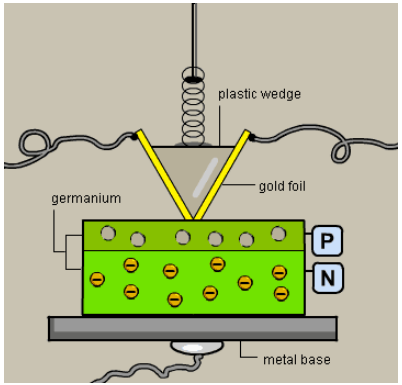
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Blue Brain Project

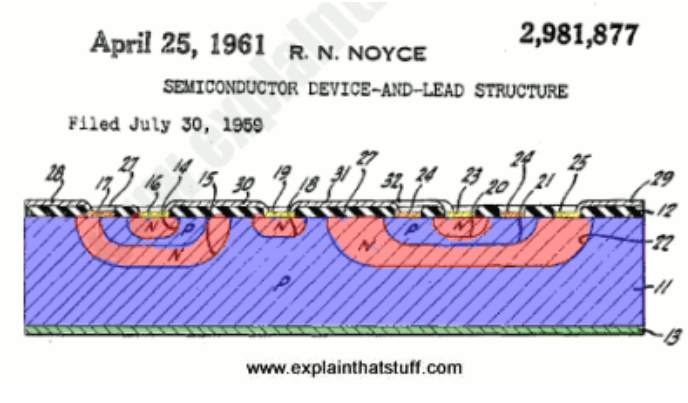
Key Enablers

Transistor - 1948

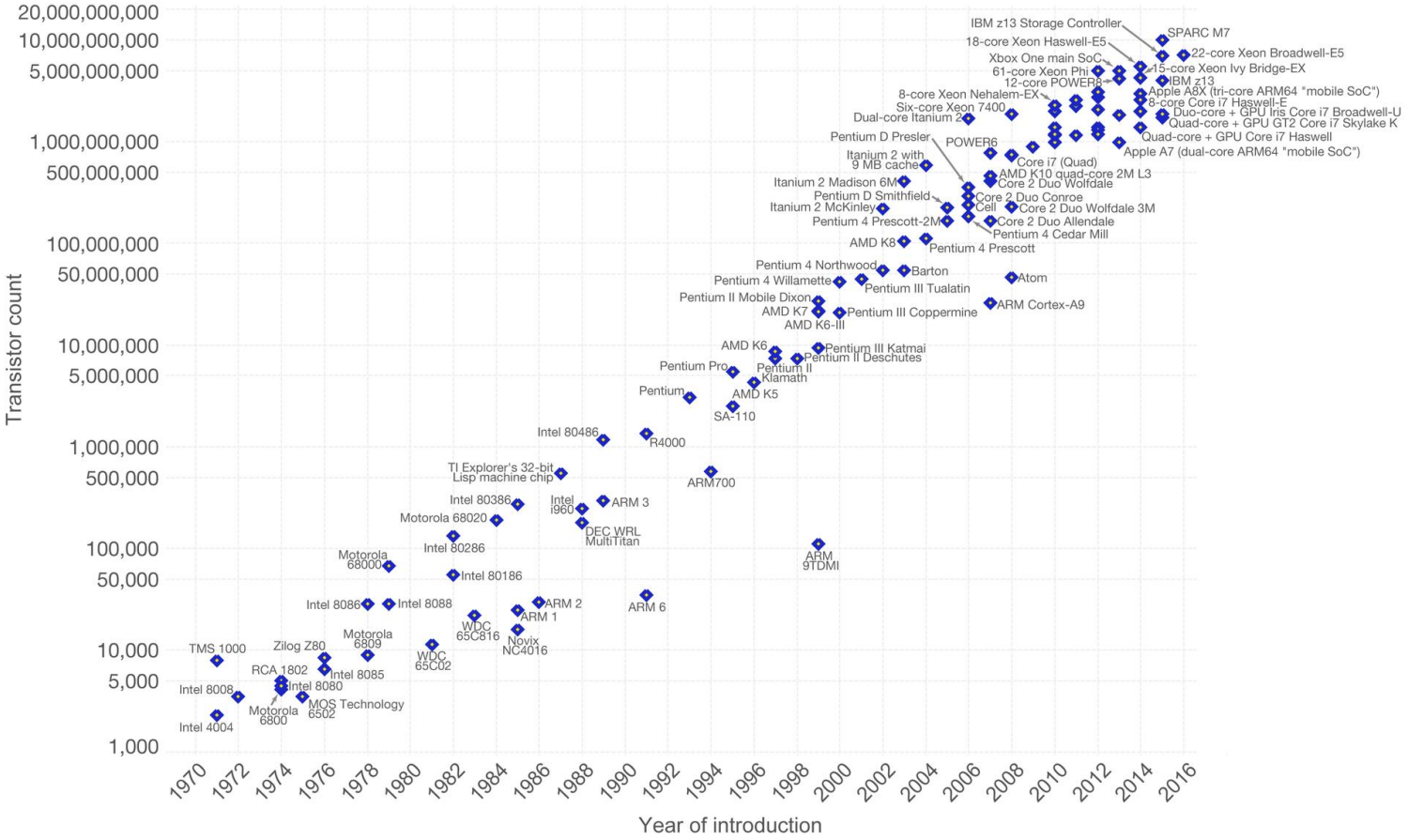




Integrated Circuit - 1958



Moore's Law

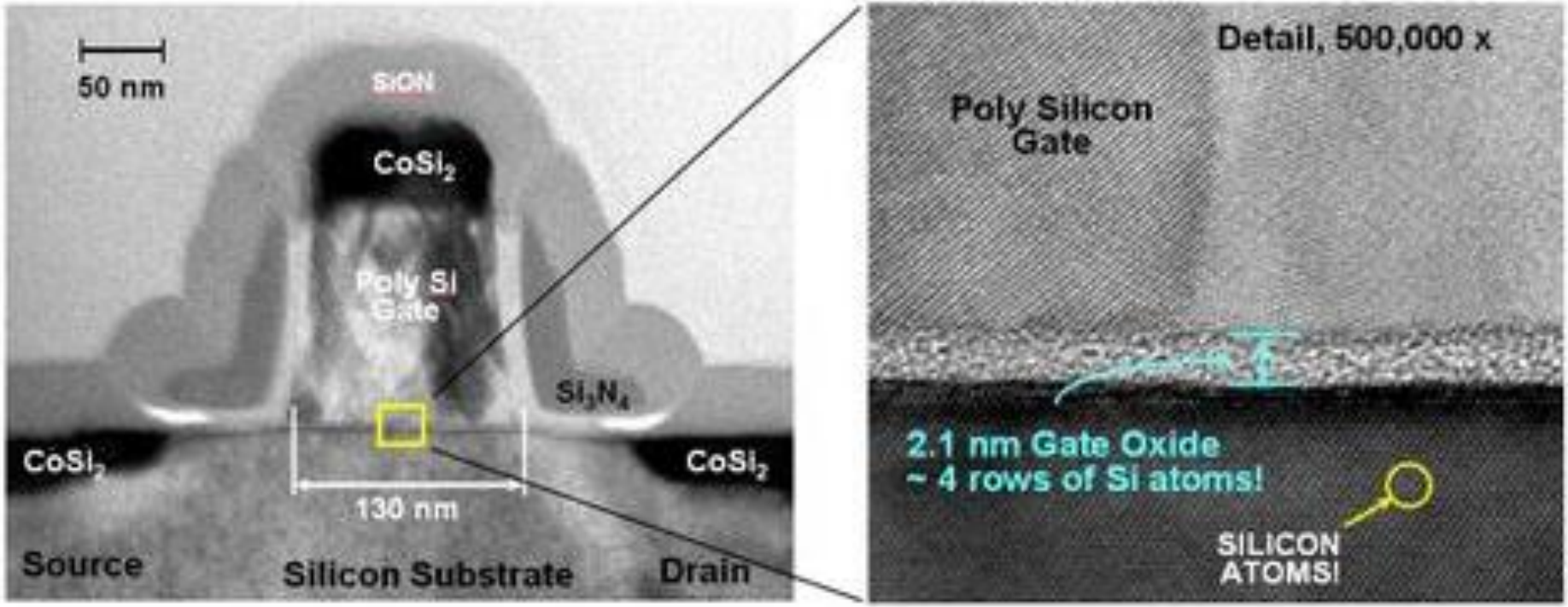


Data source: Wikipedia (https://en.wikipedia.org/wiki/Transistor_count)
The data visualization is available at OurWorldinData.org. There you find more visualizations and research on this topic.

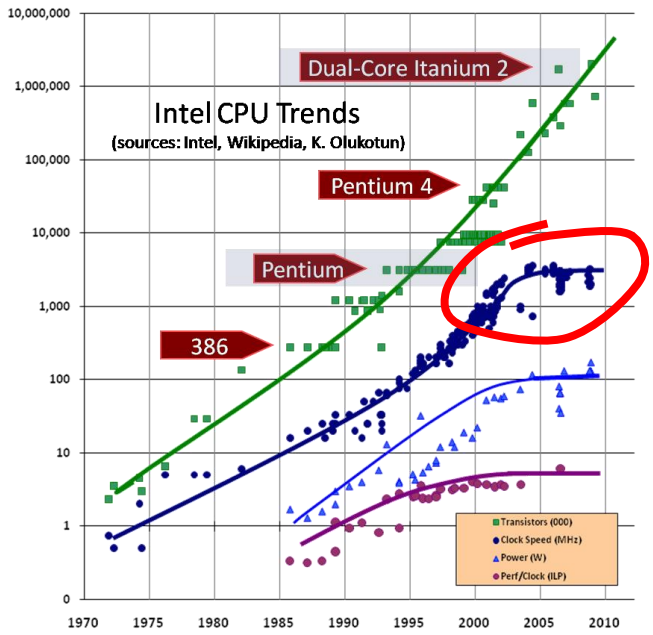
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Blue Brain Project

How does it work? – CMOS Scaling



Earlier Signs of Concern – Break Down of Strong Scaling



Source: Dr. Dobb's Journal, 30(3), 2005; update 2009 online

Shaw et al., SC2014 - Molecular Dynamics – Anton2

Anton 2: Raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer

David E. Shaw,[a] J.P. Grossman, Joseph A. Bank, Brannon Batson, J. Adam Butts, Jack C. Chao,[b] Martin M. Deneroff,[b] Ron O. Dror,[b] Amos Even, Christopher H. Fenton, Anthony Forte, Joseph Gagliardo, Gennette Gill, Brian Greskamp, C. Richard Ho,[b] Douglas J. Ierardi, Lev Iserovich, Jeffrey S. Kuskin, Richard H. Larson,[b] Timothy Layman,[b] Li-Siang Lee,[b] Adam K. Lerer, Chester Li,[b] Daniel Killebrew,[b] Kenneth M. Mackenzie, Shark Yeuk-Hai Mok,[b] Mark A. Moraes, Rolf Mueller,[b] Lawrence J. Nociolo, Jon L. Peticolas, Terry Quan, Daniel Ramot,[b] John K. Salmon, Daniele P. Scarpazza, U. Ben Schafer,[b] Naseer Siddique, Christopher W. Snyder,[b] Jochen Spengler, Ping Tak Peter Tang,[b] Michael Theobald, Horia Toma,[b] Brian Towles, Benjamin Vitale,[b] Stanley C. Wang, and Cliff Young[b]

D. E. Shaw Research, New York, NY 10036, USA

***Abstract*—Anton 2 is a second-generation special-purpose supercomputer for molecular dynamics simulations that achieves significant gains in performance, programmability, and capacity compared to its predecessor, Anton 1. The architecture of Anton 2 is tailored for fine-grained event-driven operation, which improves performance by increasing the overlap of computation with communication, and also allows a wider range of algorithms to run efficiently, enabling many new software-based optimizations. A 512-node Anton 2 machine, currently in operation, is up to ten times faster than Anton 1 with the same number of nodes, greatly expanding the reach of all-atom biomolecular simulations. Anton 2 is the first platform to achieve simulation rates of multiple microseconds of physical time per day for systems with millions of atoms. Demonstrating strong scaling, the machine simulates a standard 23,558-atom benchmark system at a rate of 85 μs/day—180 times faster than any commodity hardware platform or general-purpose supercomputer.**

	Villin	DHFR	ApoA1	ATPase	STMV	Ribosome
# atoms	13,773	23,558	92,224	327,506	1 M	2.2 M
Anton 2 (μs/day)	85.8	85.8	59.4	28.2	9.5	3.6
Anton 1 (μs/day)	19.7	18.9	7.5	—	—	—
General-purpose hardware (μs/day)	1.1[a]	0.471[b] 0.235[c] 0.144[d] 0.108[e] 0.100[f]	0.289[b] 0.127[g] 0.076[h]	0.039[i]	0.035[j] 0.018[k]	0.171[l]

→ O(100x) faster than general-purpose computers

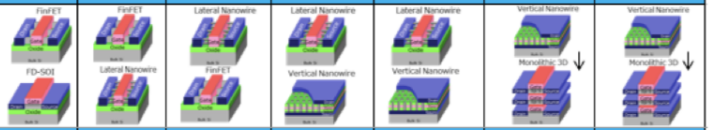
The End of Moore as we Know it for Weak Scaling?!

INTERNATIONAL ROADMAP FOR DEVICES AND SYSTEMS

2017 EDITION

- More Moore
  - "Ground rule scaling is expected to slow down and saturate around 2027"
  - "Transition to 3D integration and use of beyond CMOS devices for complementary System-on-Chip (SoC) functions are projected after 2027"
  - "technological solutions will assure continuation of Moore's Law for an additional 10−15 years"
  - BUT also: "Die cost reduction has been enabled so far by concurrent scaling of poly pitch, metal pitch, and cell height scaling. This [will likely] continue until 2024"

- Beyond CMOS
  - Emerging Architectures
  - Emerging Devices/Processes
  - Emerging Materials

YEAR OF PRODUCTION	2017	2019	2021	2024	2027	2030	2033
	P54M36	P48M28	P42M24	P36M21	P32M14	P32M14T2	P32M14T4
Logic industry "Node Range" Labeling (nm)	"10"	"7"	"5"	"3"	"2.1"	"1.5"	"1.0"
IDM-Foundry node labeling	i10-f7	i7-f5	i5-f3	i3-f2.1	i2.1-f1.5	i1.5-f1.0	i1.0-f0.7
Logic device structure options	finFET FDSOI	finFET LGAA	LGAA finFET	LGAA VGAA	LGAA VGAA	VGAA, LGAA 3DVLSI	VGAA, LGAA 3DVLSI
Logic device mainstream device	finFET	finFET	LGAA	LGAA	LGAA	VGAA	VGAA
DEVICE STRUCTURES							
							
LOGIC DEVICE GROUND RULES							
MPU/SoC Metal1x ½ Pitch (nm)[1,2]	18.0	14.0	12.0	10.5	7.0	7.0	7.0
MPU/SoC Metal0/1 ½ Pitch (nm)	18.0	14.0	12.0	10.5	7.0	7.0	7.0
Contacted poly half pitch (nm)	27.0	24.0	21.0	18.0	16.0	16.0	16.0
L_g: Physical Gate Length for HP Logic (nm) [3]	20	18	16	14	12	12	12
L_g: Physical Gate Length for LP Logic (nm)	22	20	18	16	14	14	14
Channel overlap ratio - two-sided	0.80	0.80	0.80	0.80	0.80	0.80	0.80
Spacer width (nm)	8	7	6	5	5	5	5
Contact CD (nm) - finFET, LGAA	18	16	14	12	10		
Contact CD (nm) - VGAA						12	12

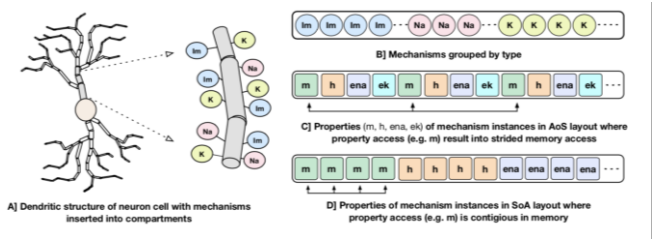
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Blue Brain Project

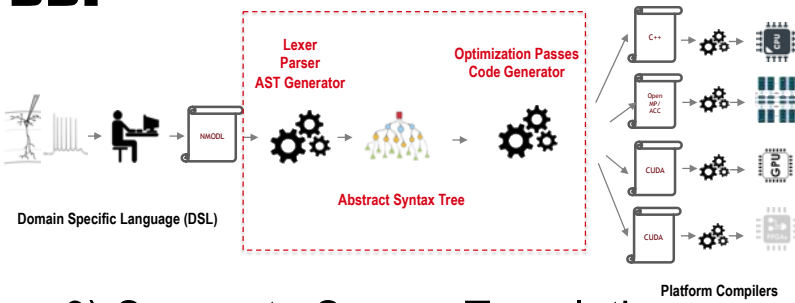
What shall we do?

- There is still substantial room for improvement in software and algorithms
  - We will have to work more on software and be ready for architectural changes
- Specialized Chips
  - GPUs, TPUs, NPUs, Spatial Accelerators etc.
- New computing paradigms: Neuromorphic, Quantum

Examples of Software Improvements in BBP



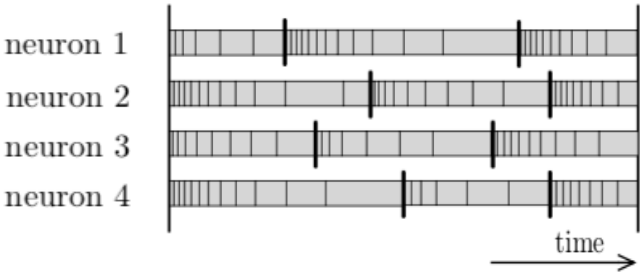
1) Classical Optimization and Scaling



3) Source to Source Translation

		runtime complexity	memory complexity
detailed COBA	ion channels	$\frac{T}{\Delta t} N_{i/n} \left(1 + N_{v/i}\right)$	$N_{i/n} \left(N_{p/i} + N_{v/i}\right)$
	membrane potential	$\frac{T}{\Delta t} N_{c/n}$	$N_{c/n} \left(N_{p/c} + N_{v/c}\right)$
	synapses	$\frac{T}{\Delta t} K \left(1 + N_{v/s}\right) + TfKN_{v/s}$	$K \left(N_{p/s} + N_{v/s}\right)$
point COBA	membrane potential	$\frac{T}{\Delta t}$	$N_{p/n} + N_{v/n}$
	synapses	$\frac{T}{\Delta t} K \left(1 + N_{v/s}\right) + TfKN_{v/s}$	$K \left(N_{p/s} + N_{v/s}\right)$
point CUBA	membrane potential	$\frac{T}{\Delta t}$	$N_{p/n} + N_{v/n}$
	synapses	$TfKN_{v/s}$	$K \left(N_{p/s} + N_{v/s}\right)$

2) Analytical Performance Modeling



4) Asynchronous Execution

5) Rethinking the problem: e.g. using machine learning

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

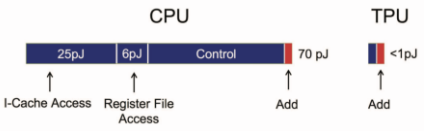
Blue Brain Project

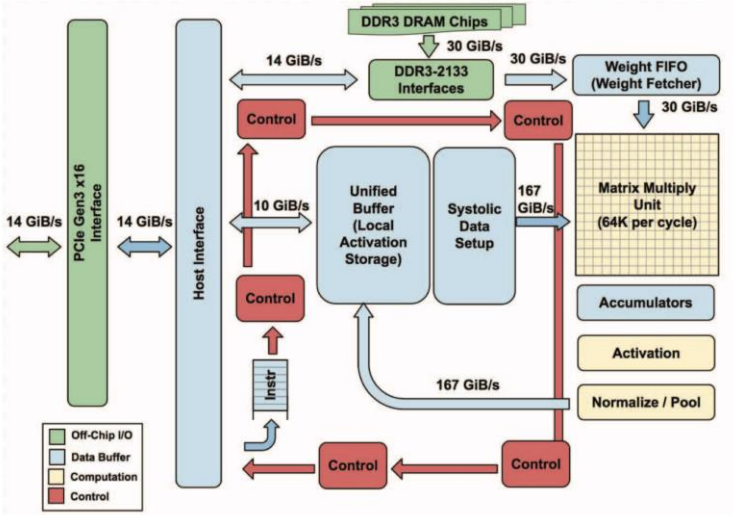
Pramod Kumbhar, Francesco Cremonesi, Bruno Magalhaes

http://arxiv.org/abs/1901.10975, *http://arxiv.org/abs/1901.05344*

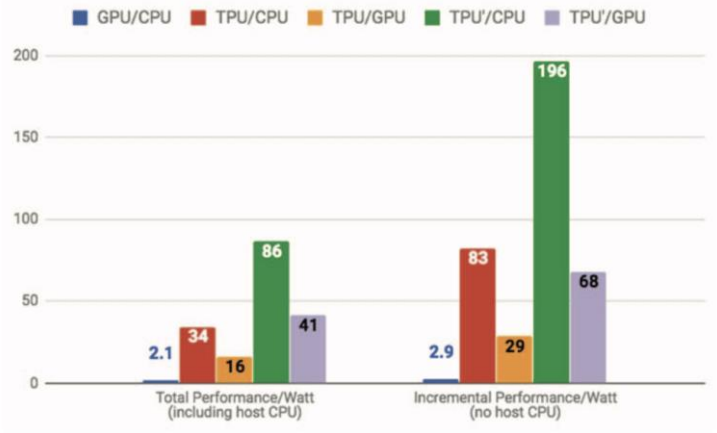
Specialized Chips

E.g. Google's TPU v1 – an ASIC specialized to calculate matrix-vector multiplications for deep neural network inference





Model	mm²	nm	MHz	Thermal Design Power (TDP) per Chip	Cores	TeraOps/s		Memory Gbyte/s	Chips/ Server	TDP per Server
						8 bit	FP32			
Haswell	662	22	2,300	145 W	18	2.6	1.3	51	2	504 W
Nvidia K80	561	28	560	150 W	13	--	2.8	160	8	1,838 W
TPU	< 331	28	700	75 W	1	92	--	34	4	861 W

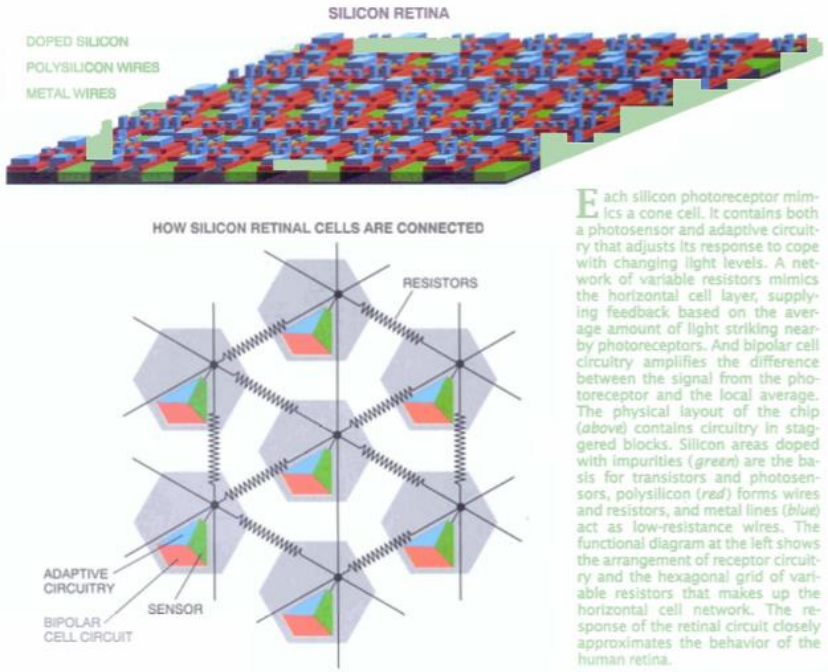


Jouppi et al., 2018

Neuromorphic Computing

"The fact that we can build devices that implement the same basic operations as those the nervous system uses leads to the inevitable conclusion that we should be able to build entire systems based on the organizing principles used by the nervous system. I will refer to these systems generically as neuromorphic systems"

Carver Mead (Caltech), 1990



1988; picture from Scientific America Article (1991)

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Blue Brain Project

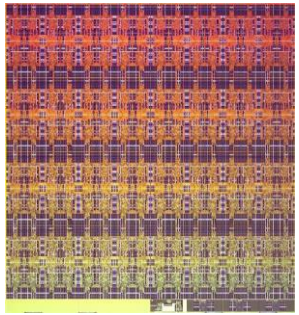
Selected Contemporary Neuromorphic Systems

SpiNNaker
Biologically Inspired Massively Parallel Architectures

IBM TrueNorth

intel Loihi

BrainScaleS



Biological realism

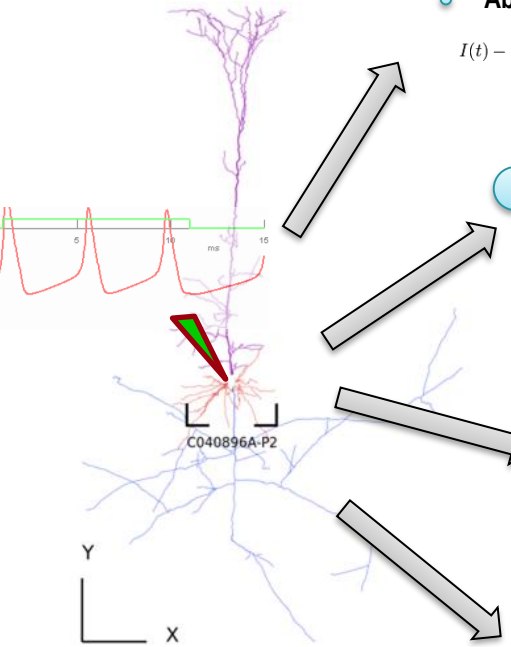
Many-core (ARM) architecture Optimized spike communication network Programmable local learning x0.01 real-time to real-time	Full-custom-digital neural circuits No local learning (TrueNorth) Programmable local learning (Loihi) Exploit economy of scale x0.01 real-time to x100 real-time	Analog neural cores Digital spike communication Biological local learning Programmable local learning x1.000 real-time

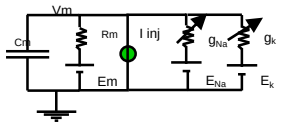
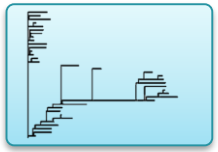
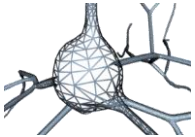
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Blue Brain Project

Also: Stanford – NeuroGrid/BrainDrop architectures
Zurich – DeltaRNN architecture

Building Blocks of Neuromorphic Hardware: Neurons

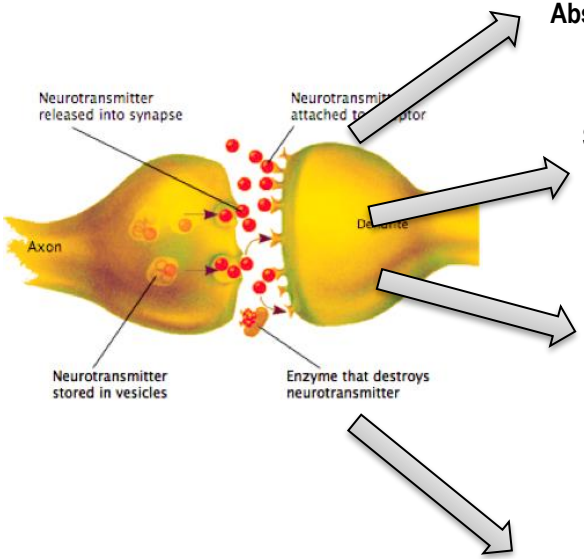


	SpiNNaker	IBM TrueNorth	intel Loihi	BrainScaleS
Abstract model: Point neuron, e.g. Leaky Integrate and Fire $I(t) - \frac{V_m(t)}{R_m} = C_m \frac{dV_m(t)}{dt}$ $f(I) = \begin{cases} 0, & I \le I_{th} \\ [t_{ref} - R_m C_m \log(1 - \frac{V_{th}}{I R_m})]^{-1}, & I > I_{th} \end{cases}$	YES numeric	YES numeric	YES numeric	YES physical
Simplified model: Single compartment and ion channel formalisms  $\frac{C_m dV_m}{dt} = \frac{E_m - V_m}{R_m} + I_{channels}$ $\frac{dm}{dt} = \alpha_m(V_m)(1-m) - \beta_m(V_m)m$ $\frac{dh}{dt} = \alpha_h(V_m)(1-h) - \beta_h(V_m)h$ $I_{channel} = m^n h g_{channel}(V_m - E_{channel})$	NO	NO	NO	NO
Cellular model: Cable and ion channel formalisms  $\frac{C_m dV_m}{dt} = \frac{E_m - V_m}{R_m} + I_{channels} + \frac{2(V_{m_{i+1}} - V_{m_i})}{R_{a_{i+1}} + R_a} + \frac{2(V_{m_{i-1}} - V_{m_i})}{R_{a_{i-1}} + R_a}$	NO	NO	COMP numeric	COMP physical
Subcellular model: Reaction-Diffusion formalism  $\dot{p}(\mathbf{x};t) = -p(\mathbf{x};t)\sum_{\mu=1}^{M} a_\mu(\mathbf{x}) + \sum_{\mu=1}^{M} p(\mathbf{x}-s_\mu;t)a_\mu(\mathbf{x}-s_\mu)$	NO	NO	NO	NO

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Blue Brain Project

Building Blocks of Neuromorphic Hardware: Synapses



Abstract model: fixed weight

$$w$$

Simplified model: exponential

$$g_{\mathrm{syn}}(t) = \sum_f \bar{g}_{\mathrm{syn}}\, \mathrm{e}^{-(t-t^{(f)})/\tau}\, \Theta\left(t - t^{(f)}\right)$$

Phenomenological: Short term plasticity

$$\begin{aligned}
\frac{dx}{dt} &= \frac{z}{\tau_{rec}} - ux\delta(t-t_{sp}) \\
\frac{dy}{dt} &= -\frac{y}{\tau_1} - ux\delta(t-t_{sp}) \\
\frac{dz}{dt} &= \frac{y}{\tau_1} - \frac{z}{\tau_{rec}} \\
\frac{du}{dt} &= \frac{u}{\tau_{\mathrm{facil}}} + U(1-u)\delta(t-t_{sp}) \\
I_{synapse}(i) &= \sum_j A_{ij} y_{ij}(t)
\end{aligned}$$

Subcellular model: Reaction-Diffusion formalism

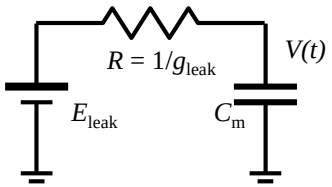
$$\dot{p}(\mathbf{x};t) = -p(\mathbf{x};t)\sum_{\mu=1}^{M} a_\mu(\mathbf{x}) + \sum_{\mu=1}^{M} p(\mathbf{x}-s_\mu;t)\, a_\mu(\mathbf{x}-s_\mu)$$

SpiNNaker	IBM TrueNorth	Intel Loihi	BrainScaleS
YES numeric	1-bit numeric	9-bit numeric	6-bit physical
YES numeric	NO	NO	6-bit physical
NO	NO	NO	NO
NO	NO	NO	NO

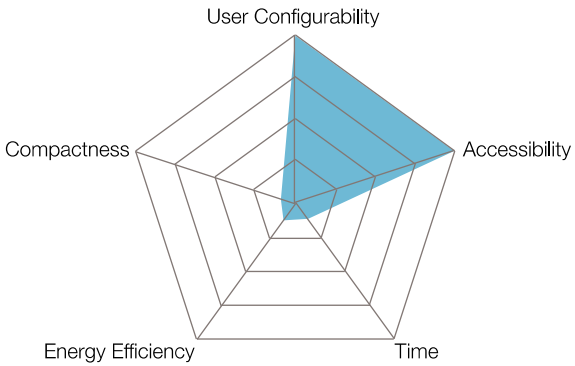
Different Design Points

$$\frac{C_m dV_m}{dt} = \frac{E_m - V_m}{R_m} + I_{channels}$$
$$\frac{d\mathrm{m}}{dt} = \alpha_{\mathrm{m}}(V_m)(1-\mathrm{m}) - \beta_{\mathrm{m}}(V_m)\mathrm{m}$$
$$\frac{d\mathrm{h}}{dt} = \alpha_{\mathrm{h}}(V_m)(1-\mathrm{h}) - \beta_{\mathrm{h}}(V_m)\mathrm{h}$$
$$I_{channel} = \mathrm{m}^n \mathrm{h} g_{channel}(V_m - E_{channel})$$

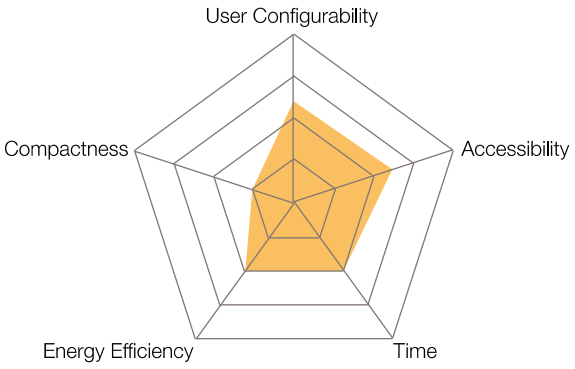
$$C_{\mathrm{m}}\frac{dV}{dt} = g_{\mathrm{leak}}(E_{\mathrm{leak}} - V)$$



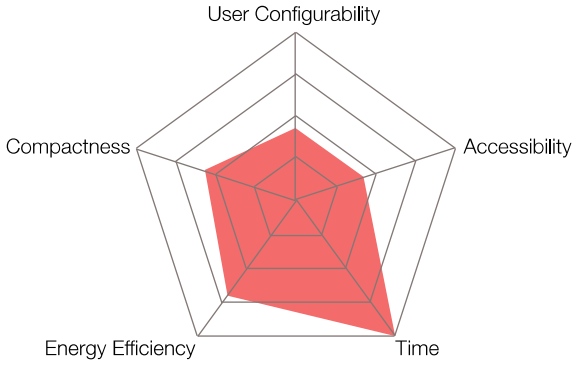
Numerical Simulation



Numerical Neuromorphic



Physical Neuromorphic



STEPS NEURON nest::

SpiNNaker IBM TrueNorth intel Loihi

BrainScaleS

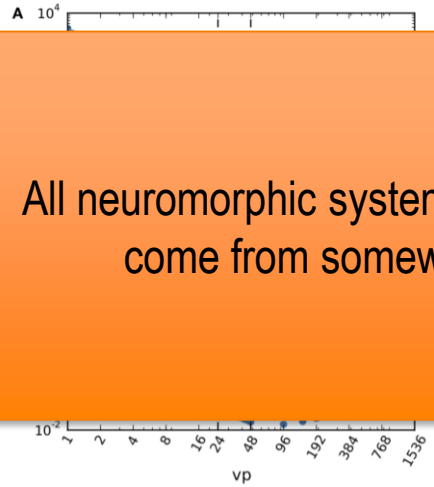
EPFL ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Blue Brain Project

Successes and Challenges of Neuromorphic Systems

- At similar time to solution numerical neuromorphic and simulation have similar power consumption; simulation can go 10x faster at cost of higher energy

- Physical neuromorphic up to 100x faster and 1000x more power efficient than simulation

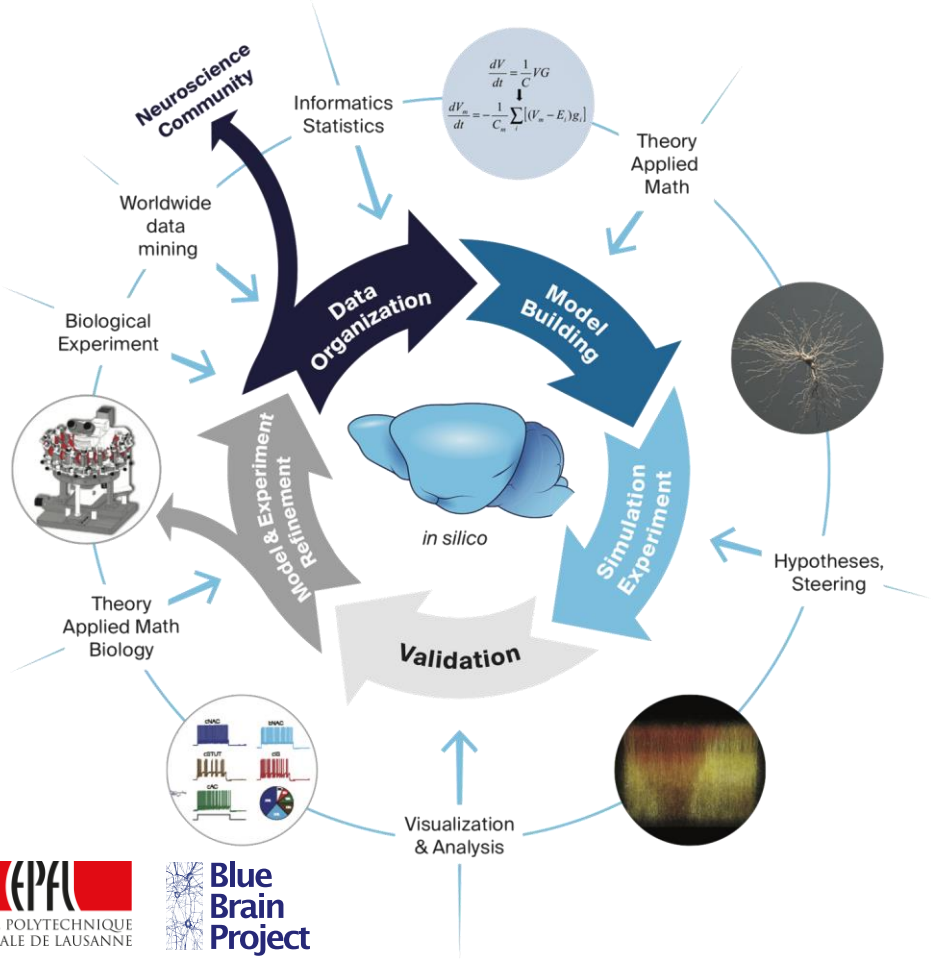


All neuromorphic systems have same challenge in common: the circuits/architecture have to come from somewhere else and in the meantime learning is supposed to fix this.

Wunderlich et al., 2019

van Albada et al., 2018

Blue Brain Project – Simulation Neuroscience



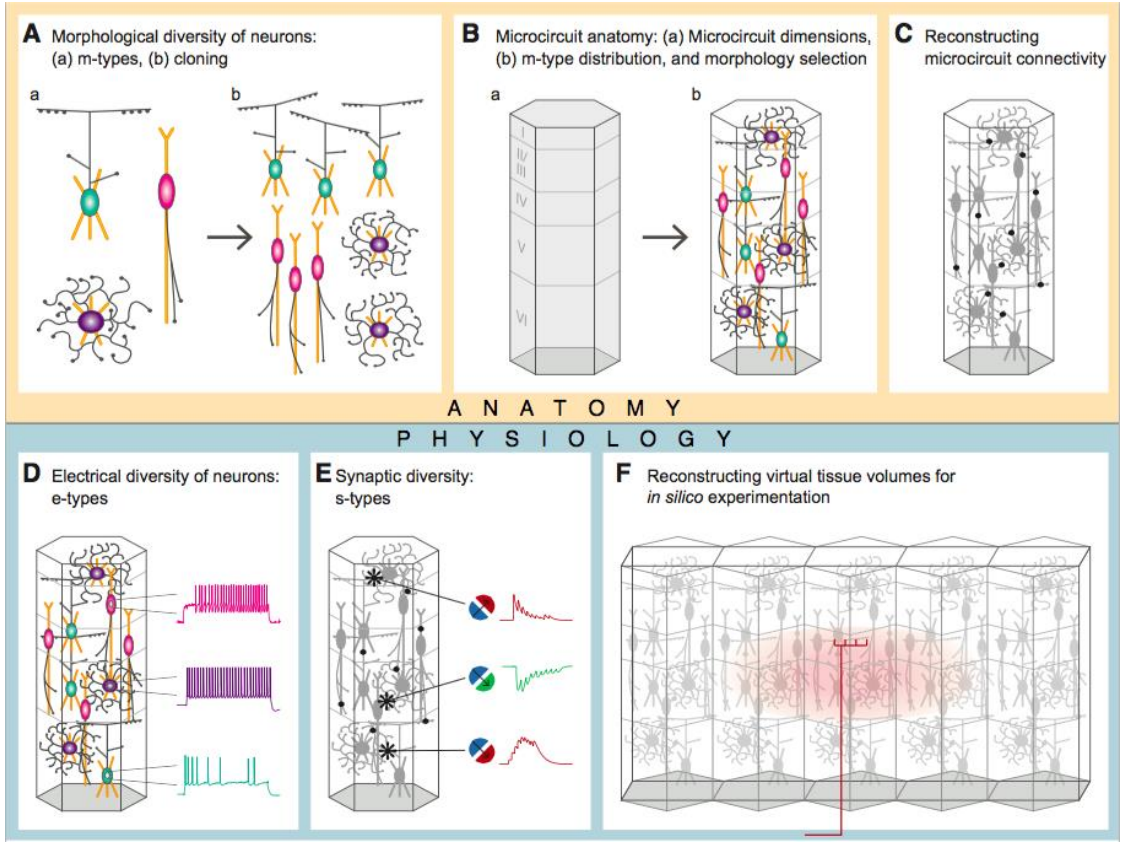
- A Swiss national infrastructure project to build digital reconstructions and simulations of the rodent, and ultimately the human brain
- Funded by the ETH Board and implemented as a center at EPFL
- Mission-driven team-science project with 130 scientists and engineers

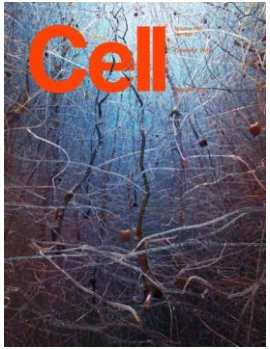
Website: https://bluebrain.epfl.ch
Public Resources: https://portal.bluebrain.epfl.ch

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Blue Brain Project

BBP Reconstruction and Simulation of Neocortical Microcircuitry





H. Markram, E. Muller, S. Ramaswamy, Michael W. Reimann, M. Abdellah, Carlos A. Sanchez, A. Ailamaki, …, Stefano M. Zaninetta, J. DeFelipe*, Sean L. Hill* I. Segev*, and **F. Schürmann* ,** *Reconstruction and Simulation of Neocortical Microcircuitry*, Cell, 2105

https://bbp.epfl.ch/nmc-portal/

- Major example of highly collaborative neuroscience: 82 authors
- Shows that detailed models can be built from sparse data
- Emergent behavior reproduces wide range of in vitro and in vivo experiments
- Can ask new questions about cellular and synaptic mechanisms

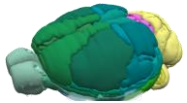
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

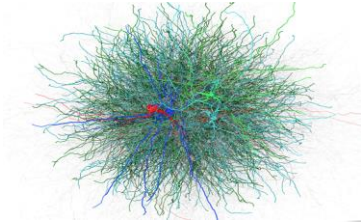
Blue Brain Project

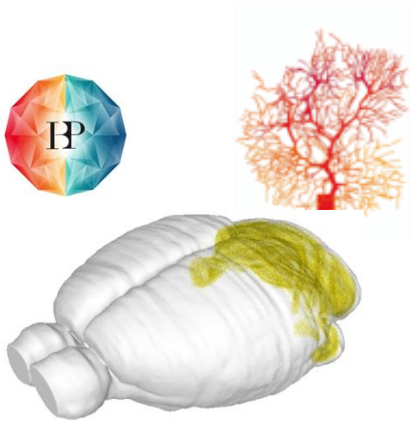
Accelerating Research towards Whole Brain

Microcircuit

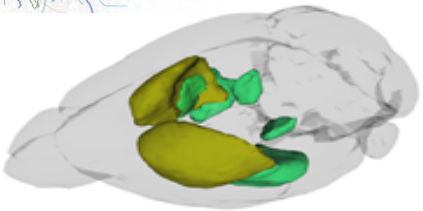
Neocortex



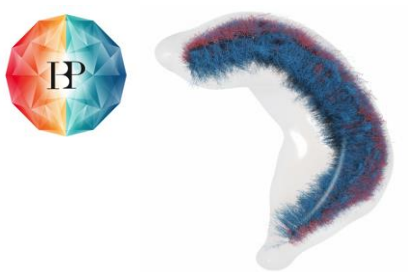




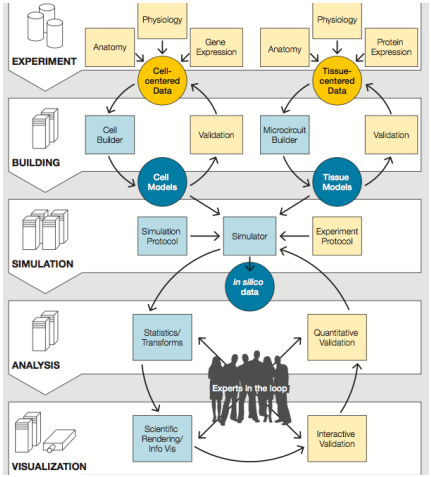
Rodent Cerebellum



Rodent Basal Ganglia



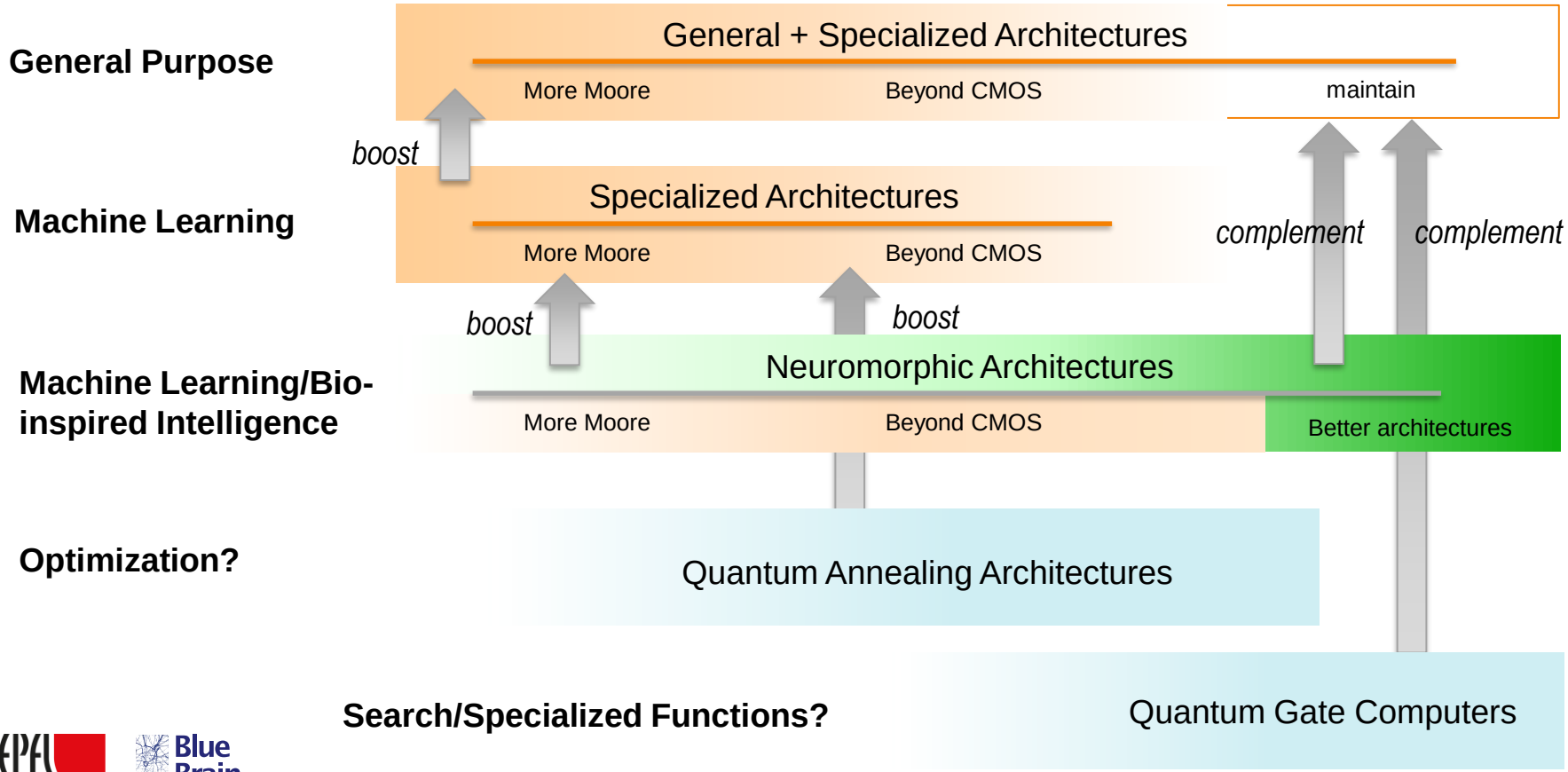
Rodent Hippocampus CA1



Blue Brain Reconstruction & Simulation Workflows

Human Neurons

Neuroscience and the Future of Computing



ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Blue Brain Project

Acknowledgments

Laboratory of Neurosimulation Technology
Francesco Cremonesi – Performance Modeling
Pramod Kumbhar – CoreNeuron, Source-2-source
Bruno Magalhaes – Asynchronous Execution
Judit Planas – Data Intensive Computing

Collaborations
Dr. Michael Hines (Yale University) – NEURON simulator
Prof. Thomas Sterling (Indiana University) – ParalleX Execution Model/HPX
Prof. Gerhard Wellein/Dr. Georg Hager – ECM Performance Modeling

The Blue Brain Team

Fabien Delalondre - CoreNeuron
Timothey Ewart – Exponential, SIMD

Funders
ETH Board
EC Human Brain Project

Computing Resources
Blue Brain Project @ CSCS
Julich Supercomputing Center
Argonne National Laboratory

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Blue Brain Project