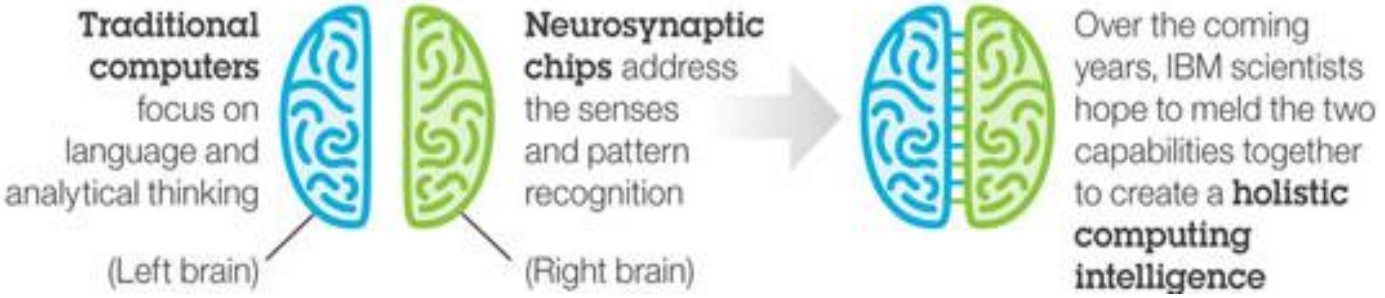
IBM's TrueNorth



Brain Computing versus Traditional Computing

- "Human-scale" simulation with 100 trillion synapses would require:
  - 96 Blue Gene/Q racks of the Sequoia supercomputer
  - ran 1,500 times slower.
  - Would hypothetically require 12GW
  - **The human brain consumes merely 20W.**
- Current microprocessors are eight orders of magnitude and four orders of magnitude hotter than the brain

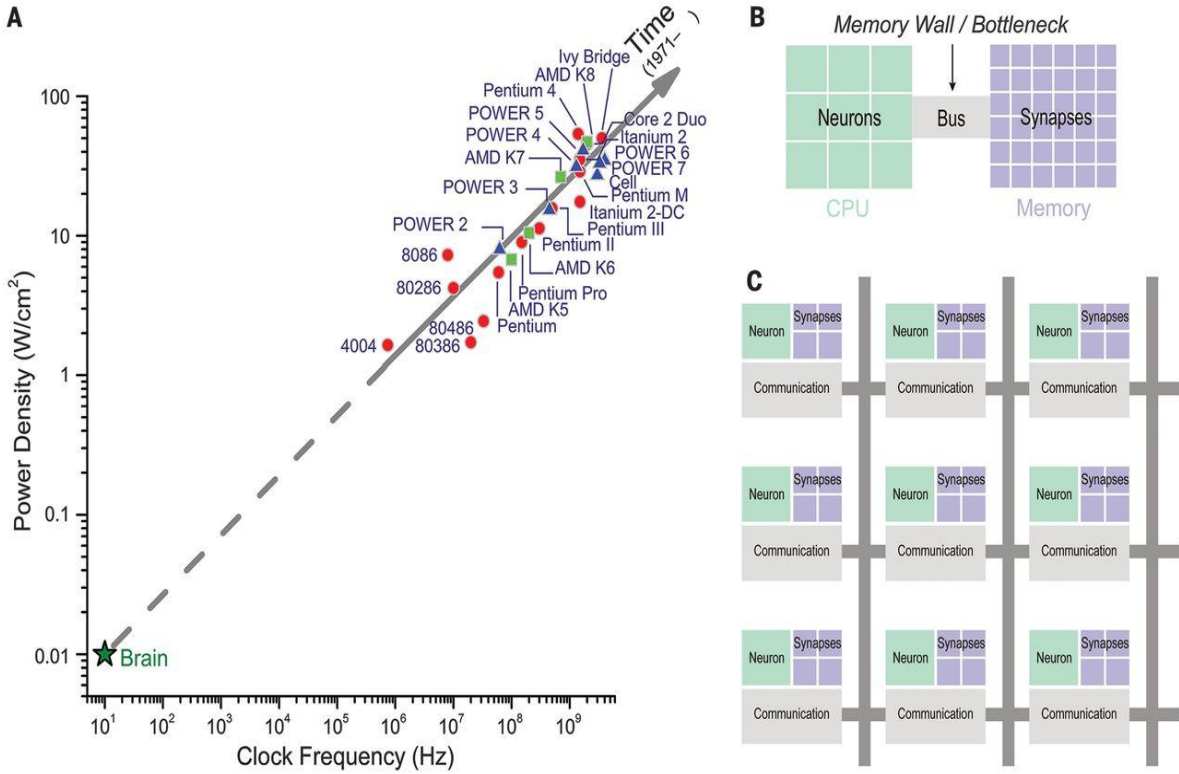


Fig. 1 Computation, communication, and memory.

TrueNorth

- World's first production-scale neuromorphic computing chip
- Non Von-Neumann
- TrueNorth Facts :-
  - 5.4 billion transistors, wired to emulate 1 million "neurons" connected via 256 million "synapses."
  - By device count, *largest IBM chip ever fabricated*, second largest (CMOS) chip in the world
  - 4,096 parallel/distributed cores, interconnected in an on-chip mesh network
  - Over 400 million bits of local on-chip memory (~100 Kb per core)
  - Design scaled down to 28nm
  - Scalable – making possible larger neural networks of several chips connected together.
  - Consumes 70 milliwatts, capable of 46 billion synaptic operations per second, per watt.
  - 20mW / cm^2 power density which comparable to cortex but is three to four orders of magnitude lower compared to 50-100W / cm^2 for a CPU

TrueNorth Eco-system

- Designed an end-to-end ecosystem
  - a new simulator, a new programming language, an integrated programming environment, new libraries, new (and old) algorithms as well as applications, and a new teaching curriculum (affectionately called, “SyNAPSE University”).

- The algorithms that run without modification on TrueNorth –
  convolutional networks
  liquid state machines
  restricted Boltzmann machines
  hidden Markov models
  support vector machines
  optical flow
  multimodal classification etc.
  200+

References

- http://www.research.ibm.com/articles/brain-chip.shtml
- http://www.sciencemag.org/content/345/6197/614.full
- http://www.sciencemag.org/content/345/6197/668.full.pdf

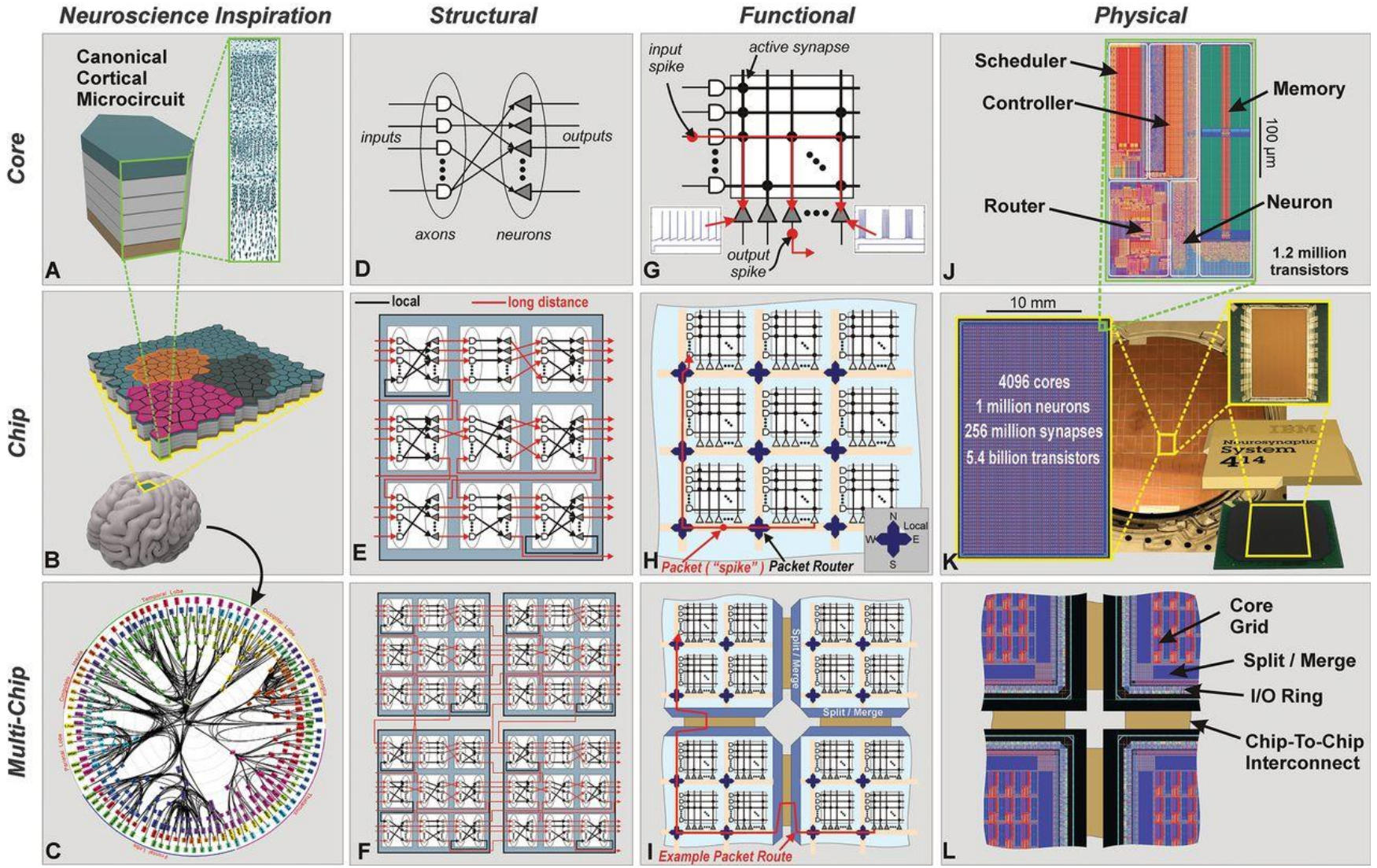


Fig. 2 TrueNorth Architecture - Panels are organized into rows at three different scales (core, chip, and multichip) and into columns at four different views (neuroscience inspiration, structural, functional, and physical).