Scaling Up of Neuromorphic Computing Systems Using 3D Wafer Scale Integration

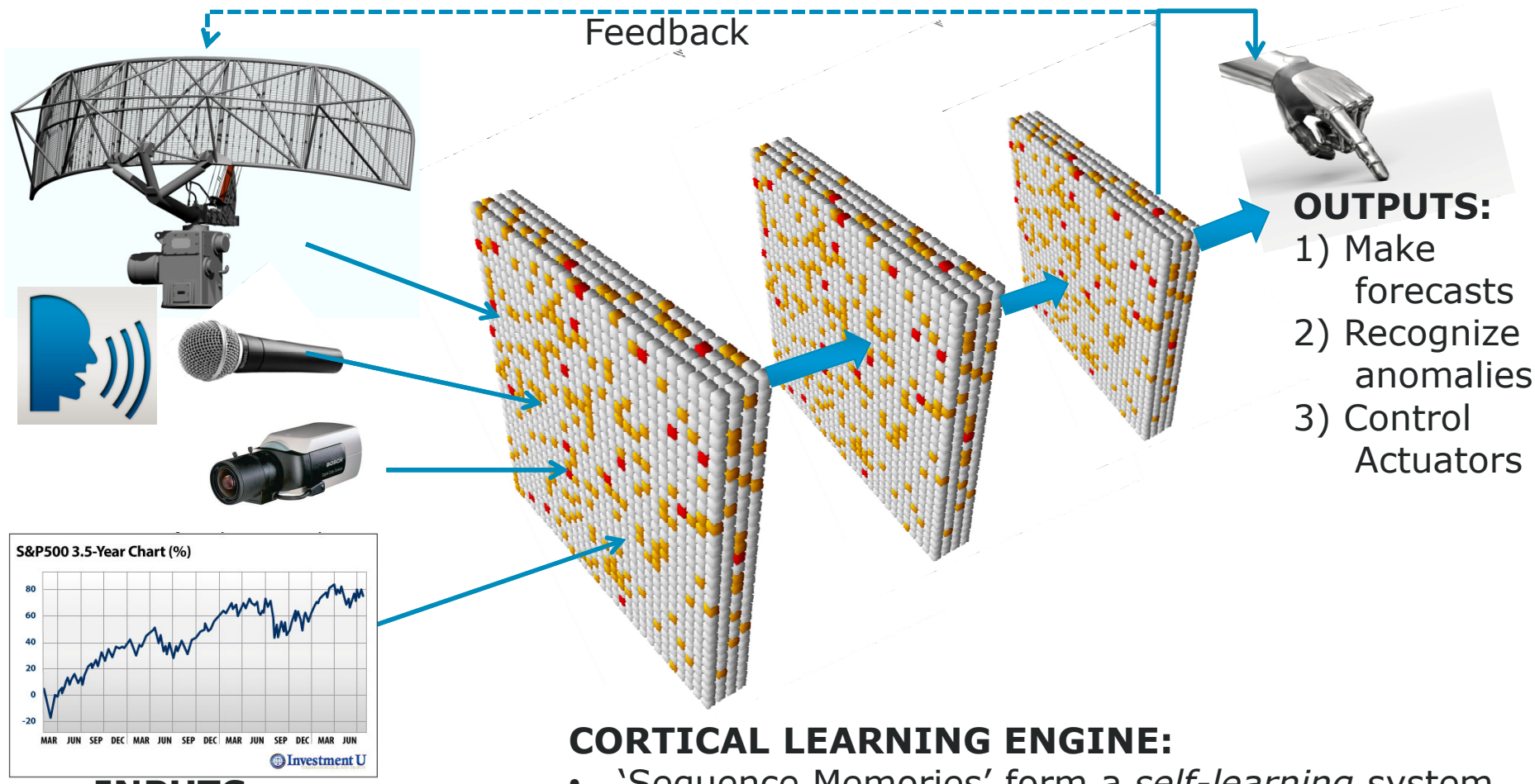
Arvind Kumar
IBM Thomas J Watson Research Center

Zhe Wan
IBM Albany Nanotech Center and UCLA

Subramanian S. Iyer
UCLA

Winfried W. Wilcke
IBM Almaden Research Center

Goal of IBM's Center for Machine Intelligence



INPUTS:
Any type of spatial-temporal data stream

OUTPUTS:
1) Make forecasts
2) Recognize anomalies
3) Control Actuators

CORTICAL LEARNING ENGINE:
- 'Sequence Memories' form a *self-learning* system
- Detect and predict patterns in the input streams
- Based on 'Hierarchical Temporal Memory' theory

Machine Learning ML vs. Machine Intelligence MI

- **ML and MI are very different beasts**

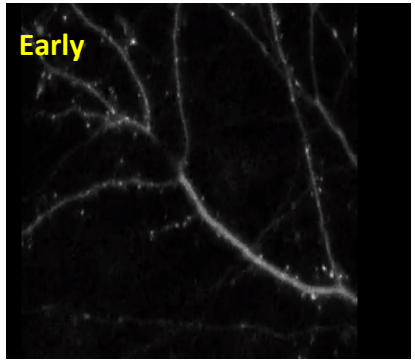
- **Machine Learning: Consists of solving a *specific* task by defining and optimizing an objective function (Yann LeCun)**
  - e.g. Deep Learning with Neural Networks
  - training is (usually) supervised from labeled datasets and distinct from testing /execution

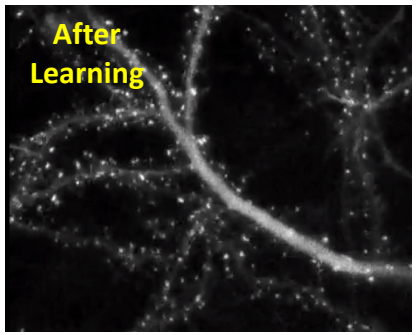
- **Machine Intelligence: Cognitive systems which *learn continuously* and often *without supervision*, are *universal*, *predict* patterns and temporal sequences and detect *anomalies*. May perform motor actions to achieve goals.**

- **Machine Learning => Machine Intelligence**

*H*ierarchical *T*emporal *M*emory (HTM)

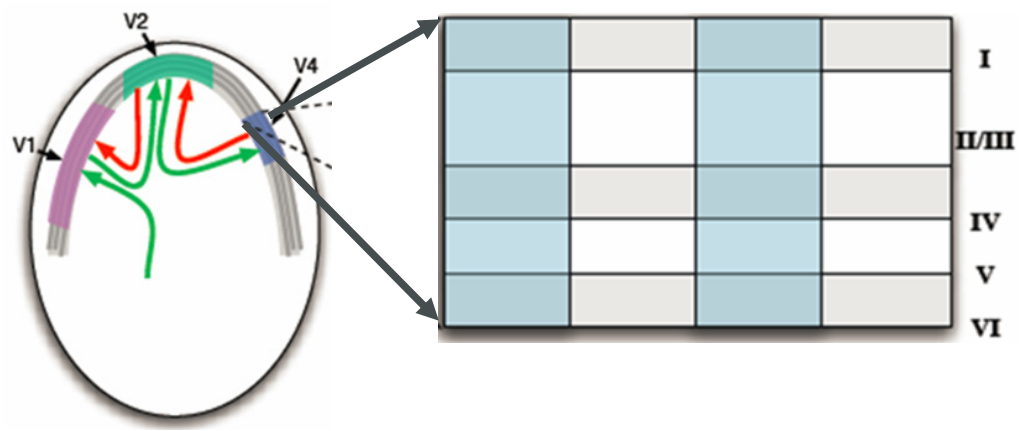
- **A system-level model of some of the structural and algorithmic behavior of the neocortex**
- **It's build around unsupervised / online learning**
  - machine intelligence, not machine learning
- **Hits a sweet spot in biological fidelity**
  - learning occurs through formation of synapses rather than via fine-grained weight changes
- **It is a rather universal model**
  - conceptually it always does the same
  - what it does depends on the sensors/actuators it is hooked up to
  - no need for new software for each new application

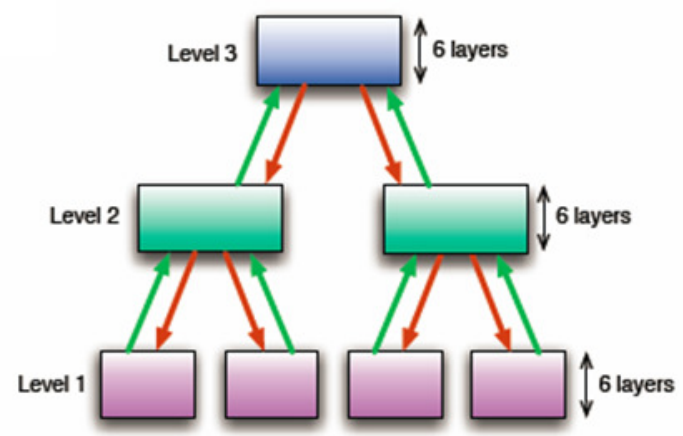


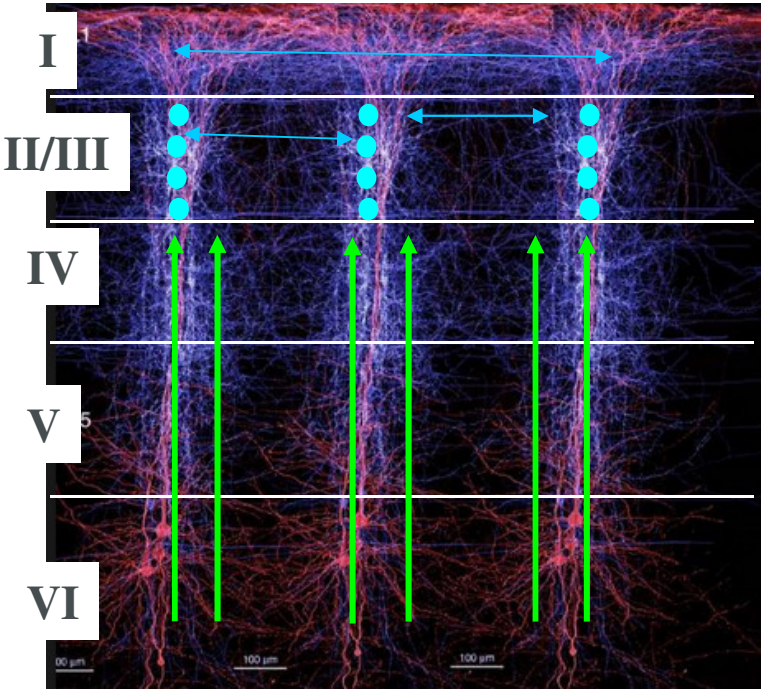


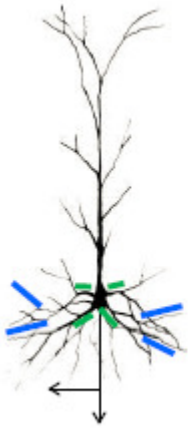
Minerbi et al. PLoS Biol 7(6), 2009

HTM Cortical Model

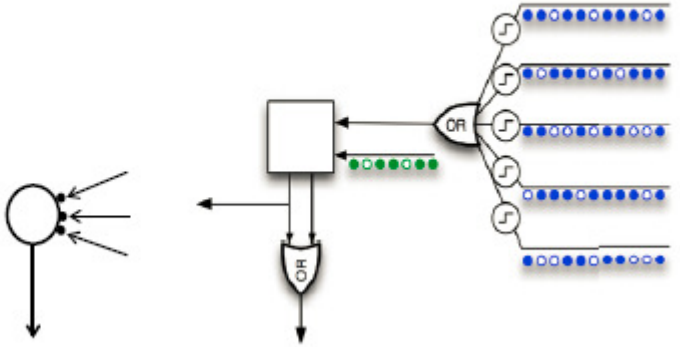








single
neuron
in layer II/III

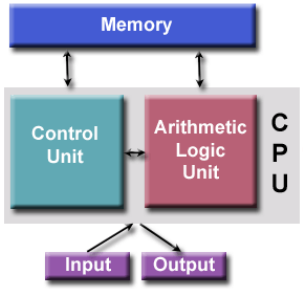


green: feedforward connections
blue: lateral connections

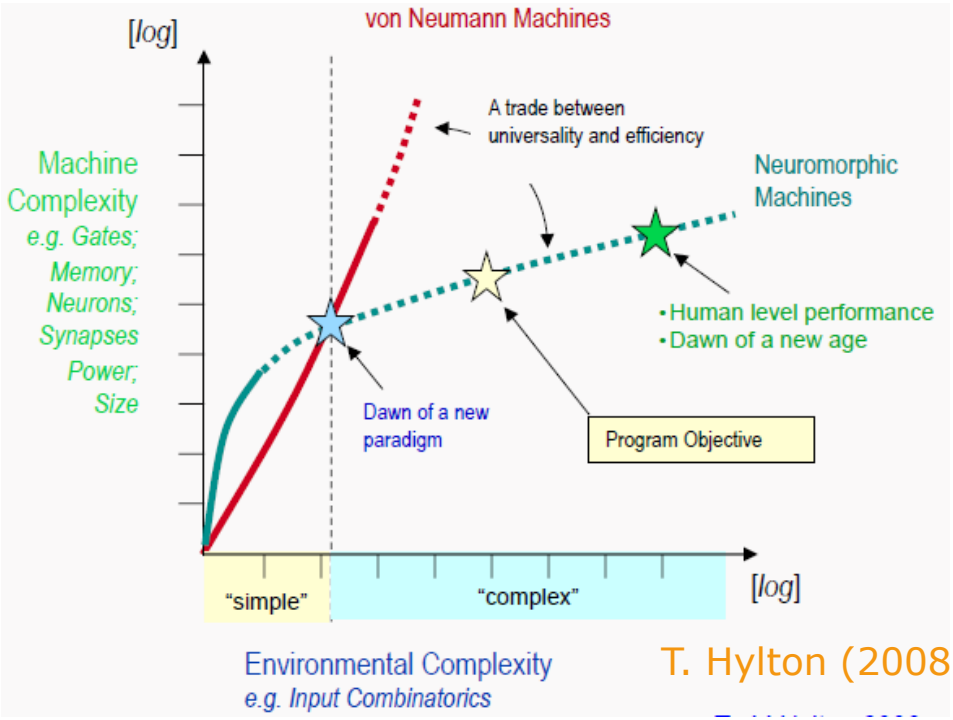
3D-WSI for Neuromorphic Computing

- **The extremely high connectivity of wafer-scale and 3D stacking is a great match for building a cortical system**
  - Performance is derived from high memory bandwidth feeding a large number of fairly simple processors
  - Very high communications performance between processors (message passing model)
- **The resilience of HTM algorithms (shown via simulations) makes wafer-scale yield problems much less of a concern**
- **Neuromorphic applications are naturally low power**

von Neumann vs Neuromorphic architecture



von Neumann:
Logic & memory separated
Centralized, sequential processing

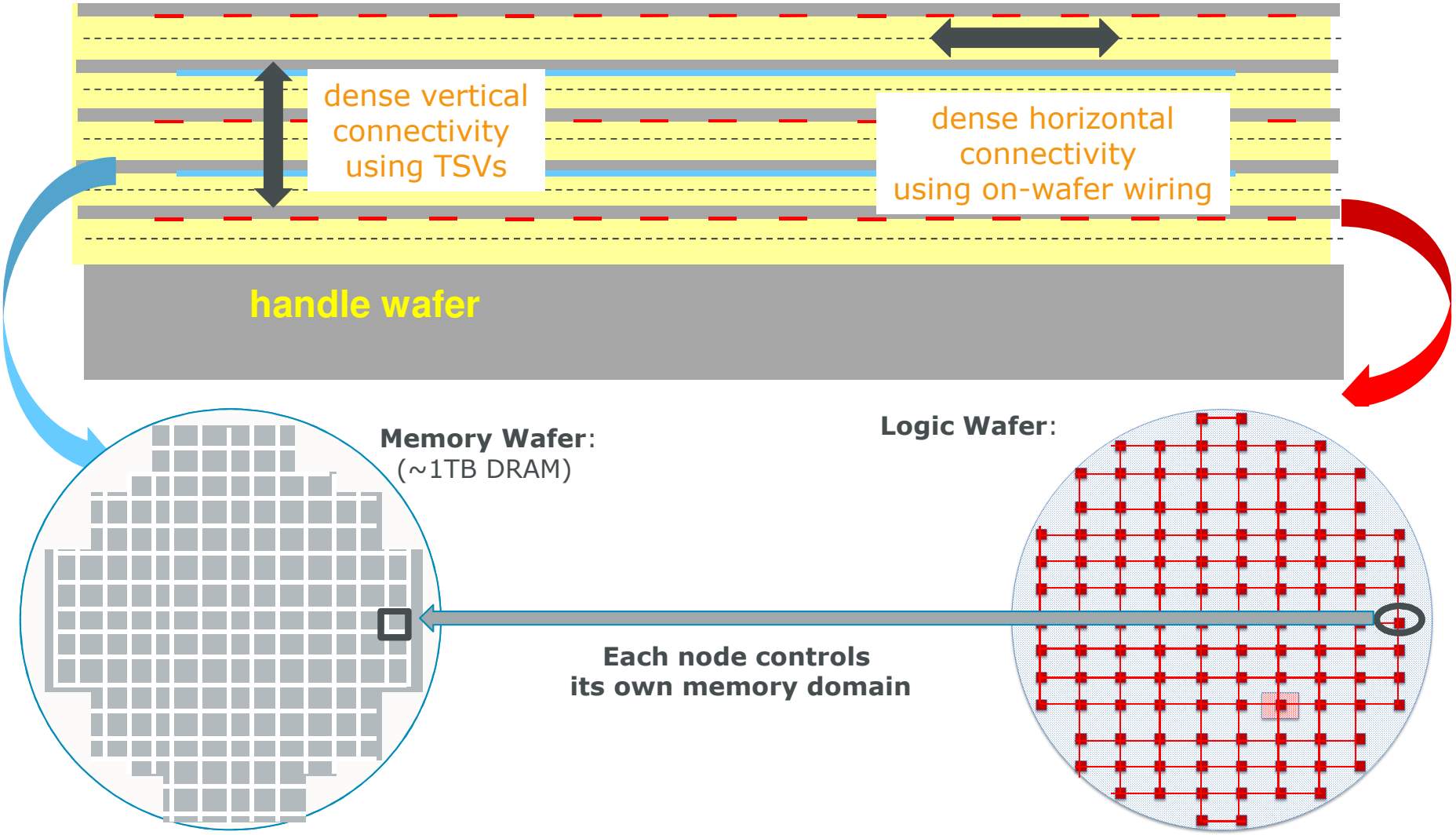


T. Hylton (2008)

Neuromorphic:
Logic & memory integrated
Distributed, parallel processing

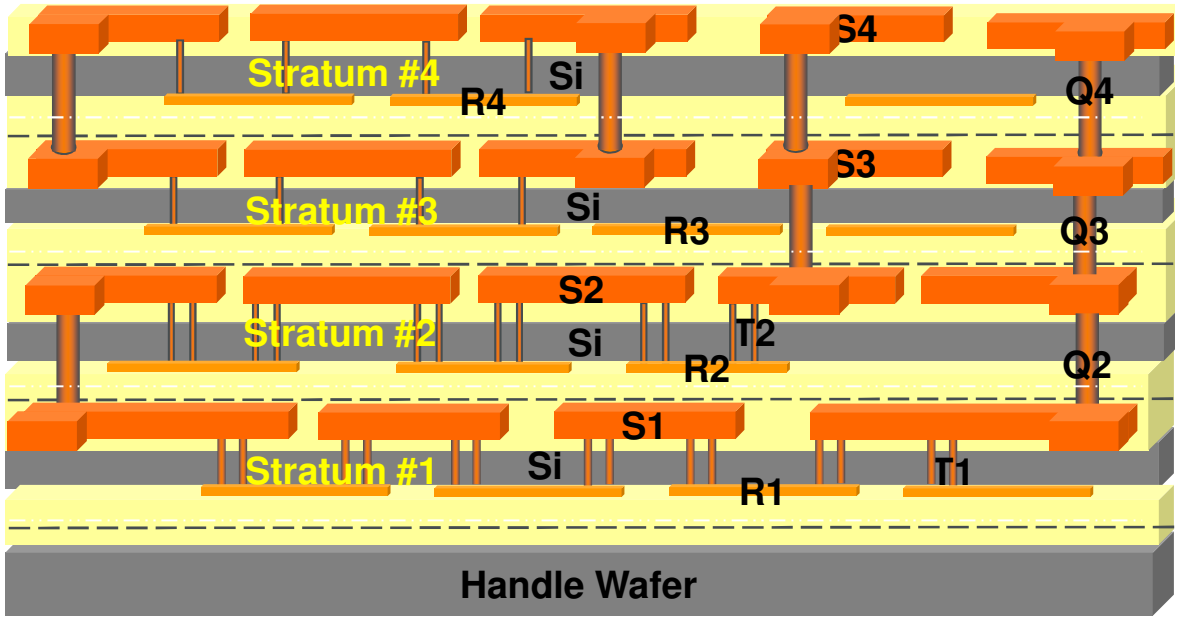
Architecture:
- Distributed, de-centralized processing
- Memory-centric rather than Compute-centric architecture

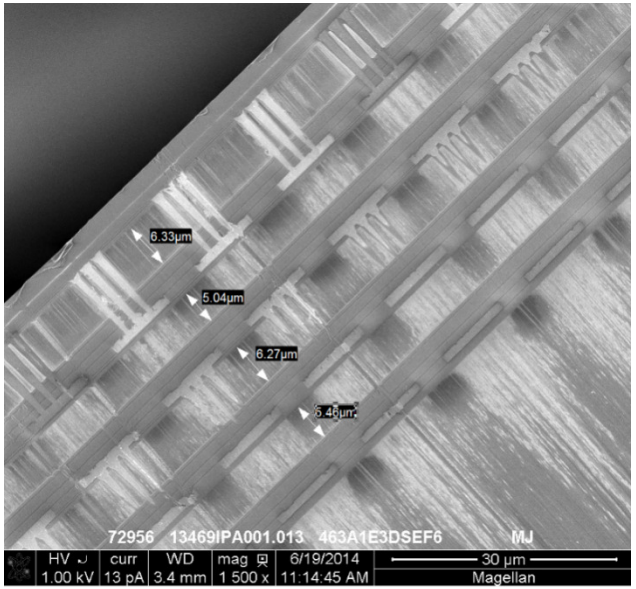
Cortical System using 3D-WSI



“Gen III” 3D Integration: Wafer-to-Wafer Stacking

IBM Albany Nanotech Center





<u>Wafer Scale Integration</u>:

Use an entire Si wafer to build a large-scale system on a wafer

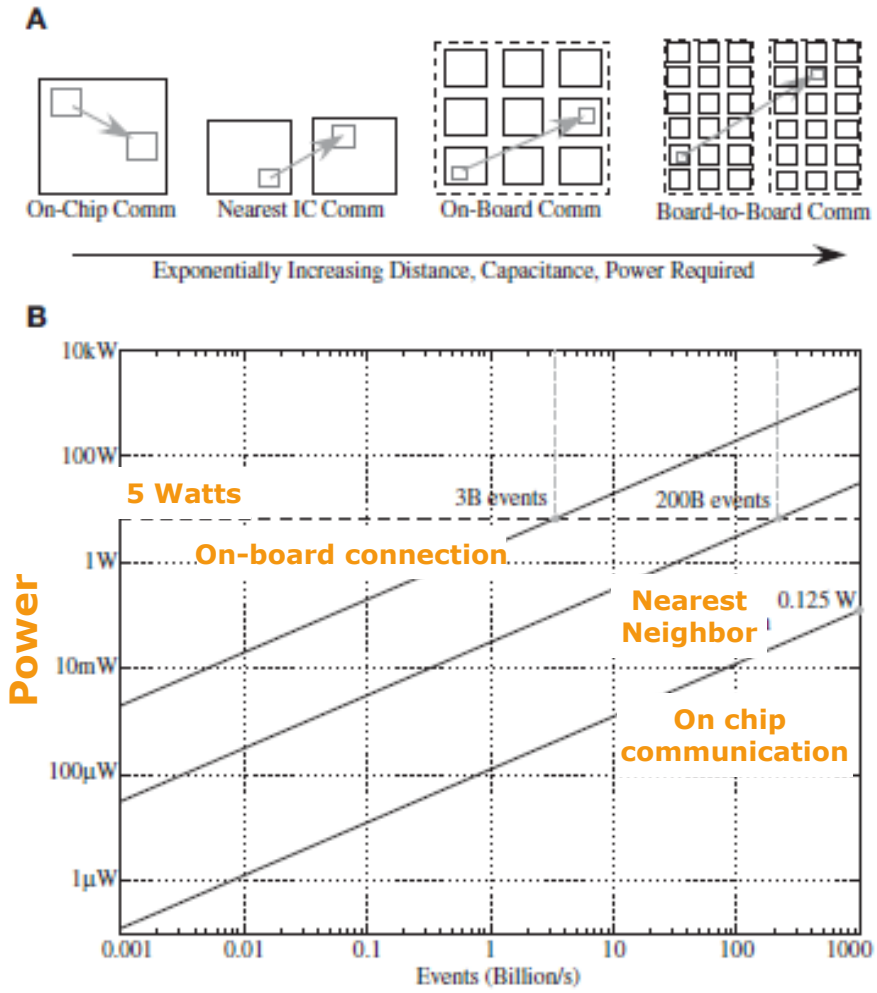
Lin et al., IEEE S3S Conf., 2014

<u>3D Wafer-to-Wafer Stacking</u>:

Use wafer bonding techniques and TSVs to connect multiple wafers in a stacked configuration

- **Supports dense inter strata connectivity**
- **Fault tolerance and repair techniques are central to design**
- **Allows volume production**

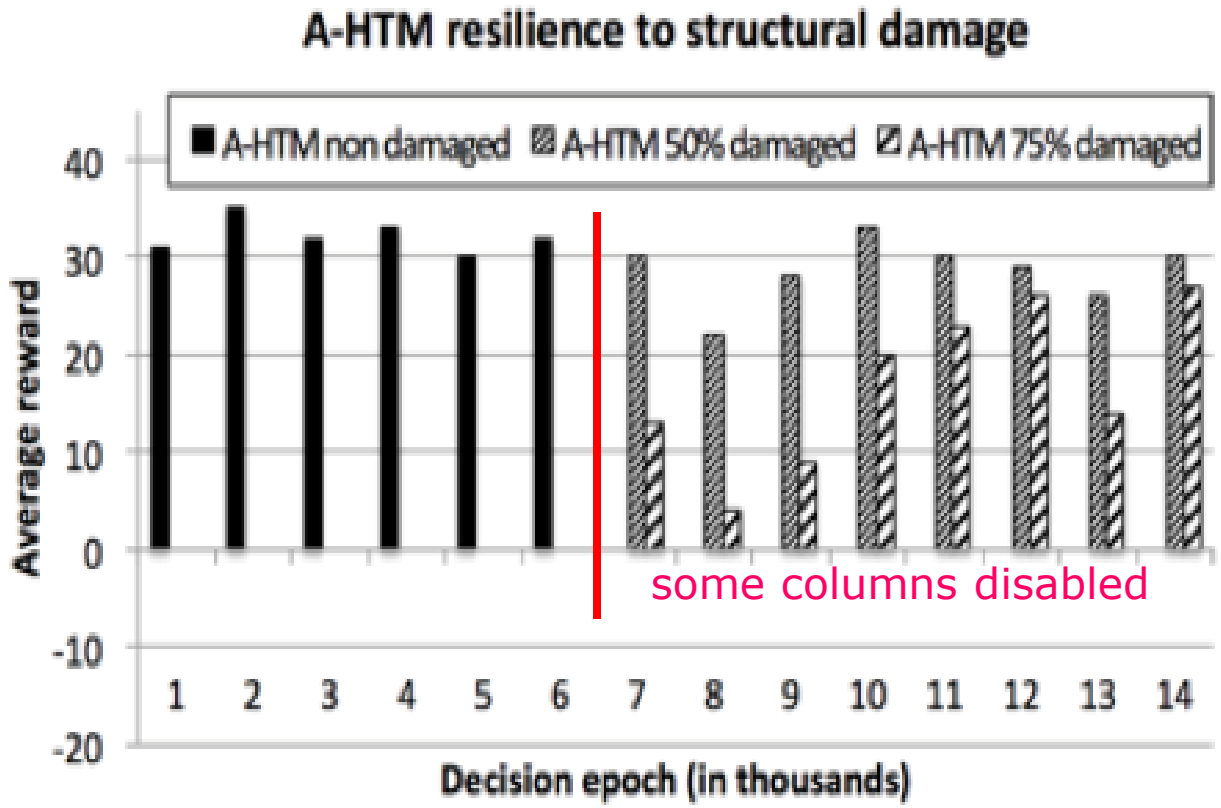
Power Cost of Communication



J. Hasler & B. Marr, Frontiers in Neurosci., Sep. 2013

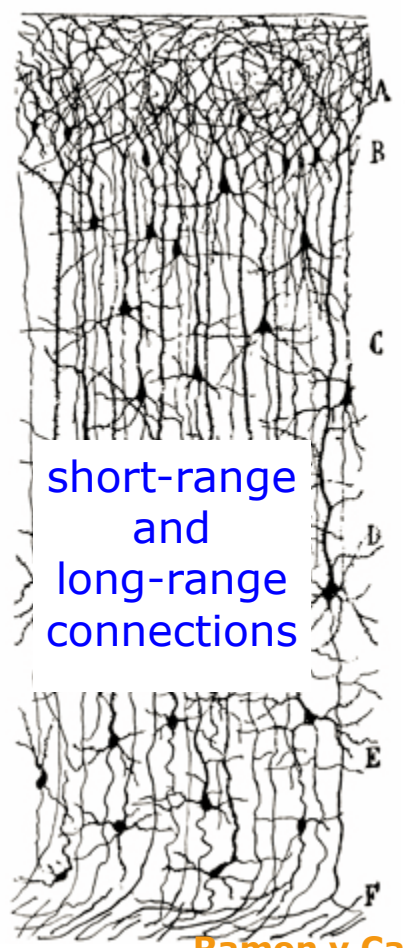
Power constraint dictates close proximity

Natural Fault Tolerance of Neuromorphic Applications

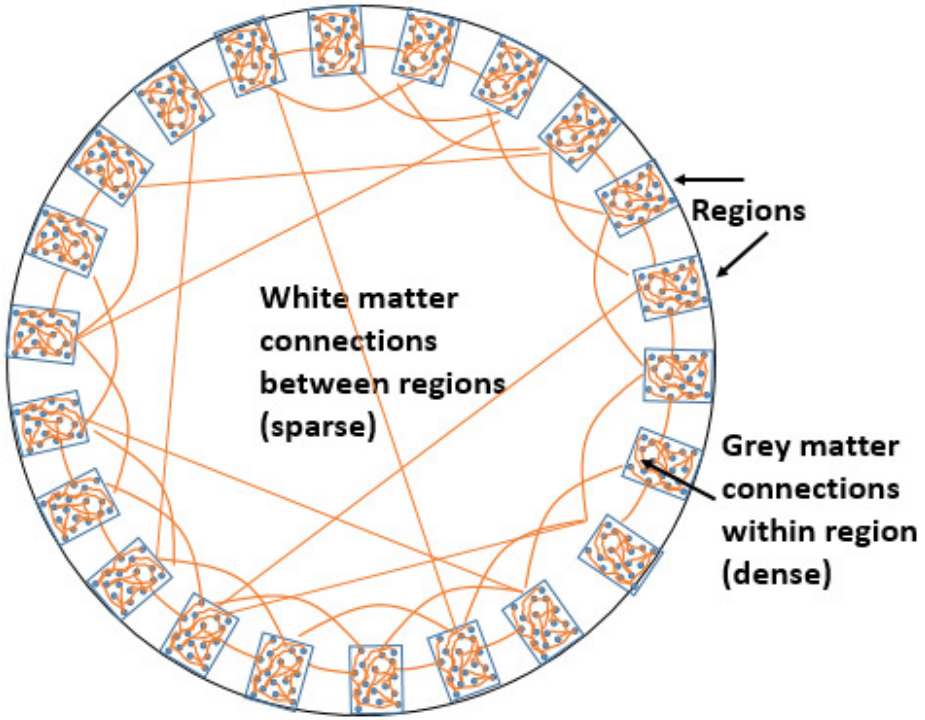


Effects of destroying 50% or 75% of columns in a-HTM (result from Janusz Marecki)

Short- and Long-range Connectivity Model



Ramon y Cajal



Conclusions from model study:

- Bandwidth is quite sufficient (10's of Gbps per processor)
- Latency is not BW-limited and can be hidden behind compute cycle
- Power is high (10's of kW) but manageable
- Memory capacity is biggest constraint for very high synaptic connectivity