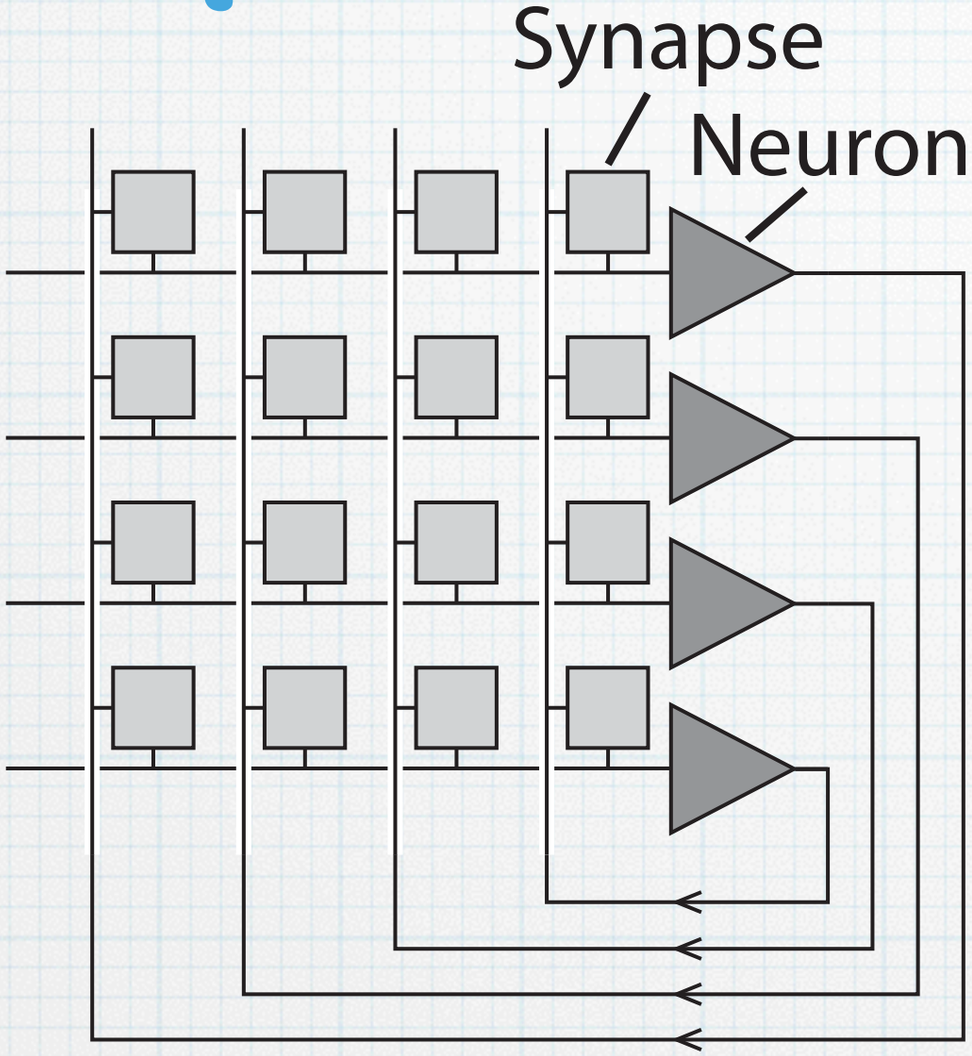
# Architectures: How do area, energy and time scale with number of neurons (N)?

Benjamin et al., 2014

# Fully Dedicated Analog (FDA)



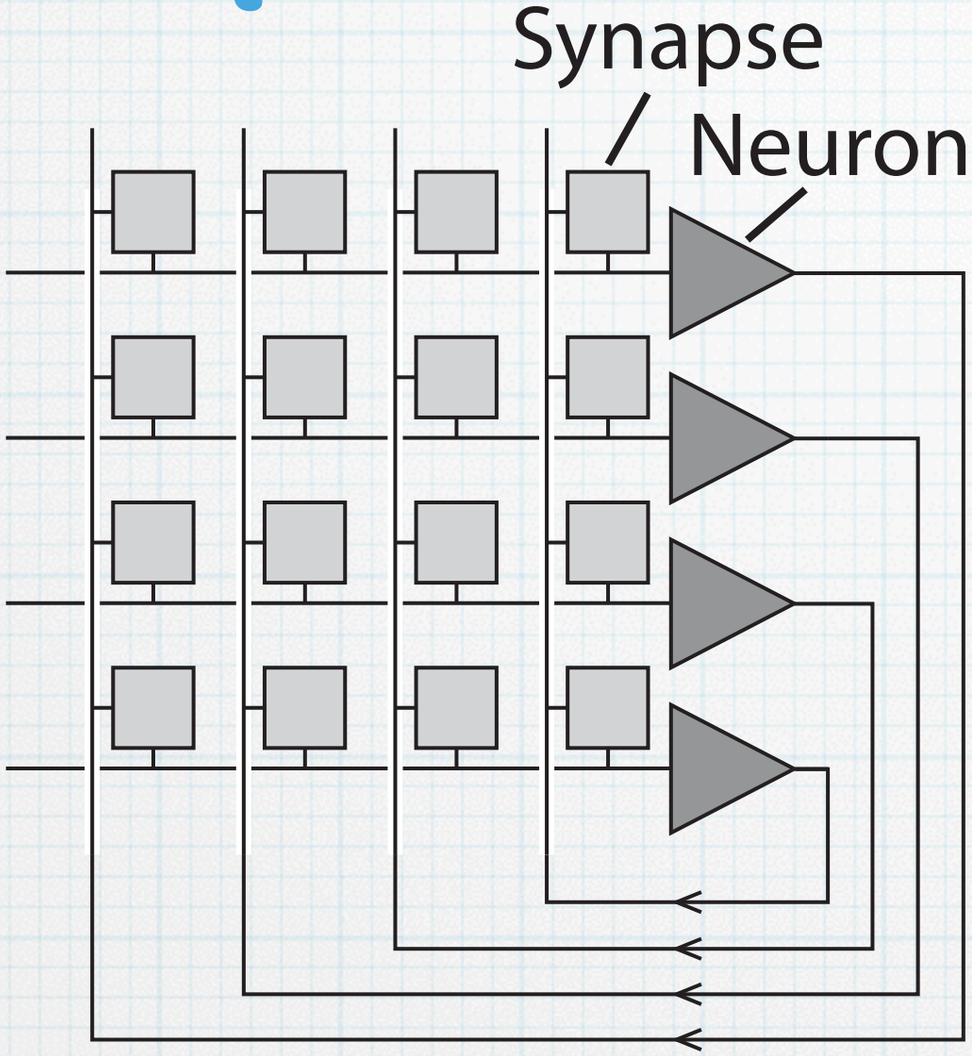

Sivilotti et al. 85
Boahen & Andreou 89

| Per Synapse | |
|---|---|
| Area | 1 |
| Energy | 1+1 |
| Time | 1/N |
| A×E×T | 1/N |

* $E = C_{axo}V_{axo}^2/N + C_{den}V_{den}V_{DD}$
* $T = (C_{axo}V_{axo}/I_{axo})/N^2$

- A synapse has area 1x1; a dendrite or axon has capacitance $C_{axo}=C_{den}=N$
- A digital axon's voltage change is $V_{axo}=V_{DD}=1$; an analog dendrite's is $V_{den}=1/N$
- An axon is driven by a current $I_{axo}=1$; all N axons may be driven in parallel

# Fully Dedicated Digital (FDD)



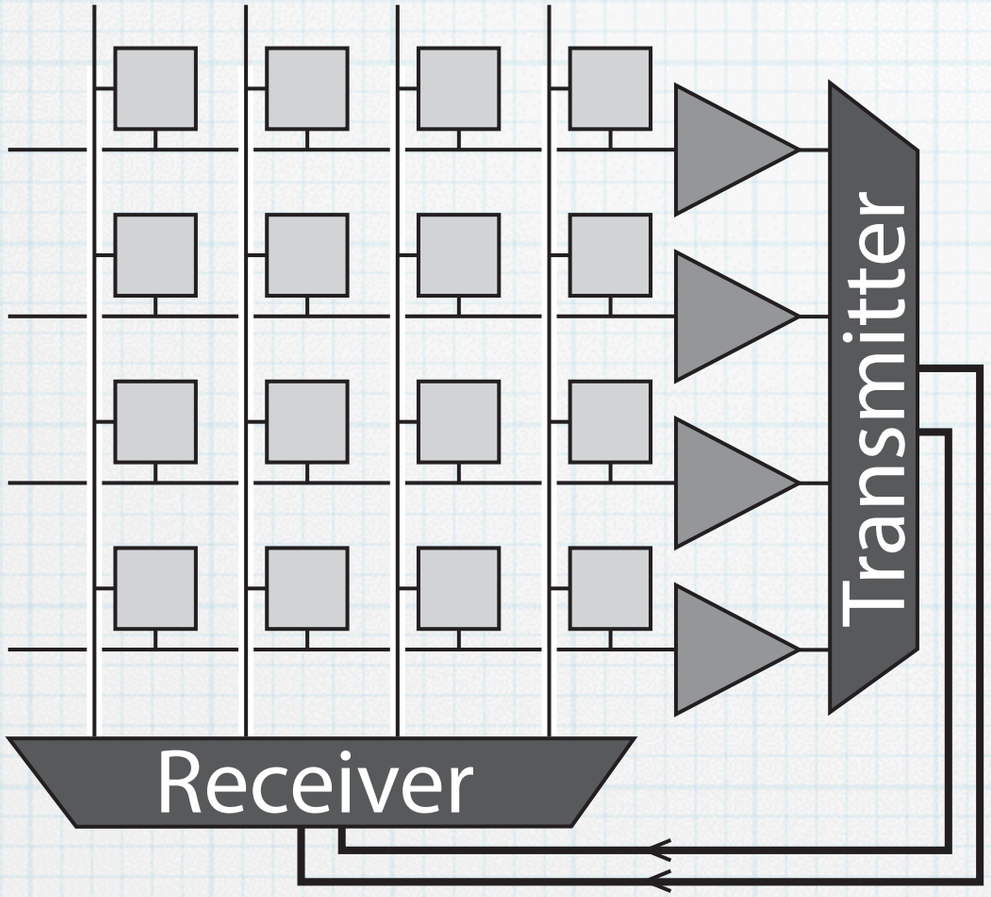

Merolla et al 11

| Per Synapse | |
|---|---|
| Area | 1 |
| Energy | 1+N |
| Time | 1 |
| A×E×T | 1+N |

* $E = C_{axo}V_{axo}^2/N + C_{den}V_{den}V_{DD}$
* $T = (C_{axo}V_{axo}/I_{axo})/N$

- A digital dendrite's voltage change is $V_{den}=1$—versus $V_{den}=1/N$ for analog
- When dendrites are digital, only one axon may be active at any time

# Shared Axon (SA)




Indiveri et al 04

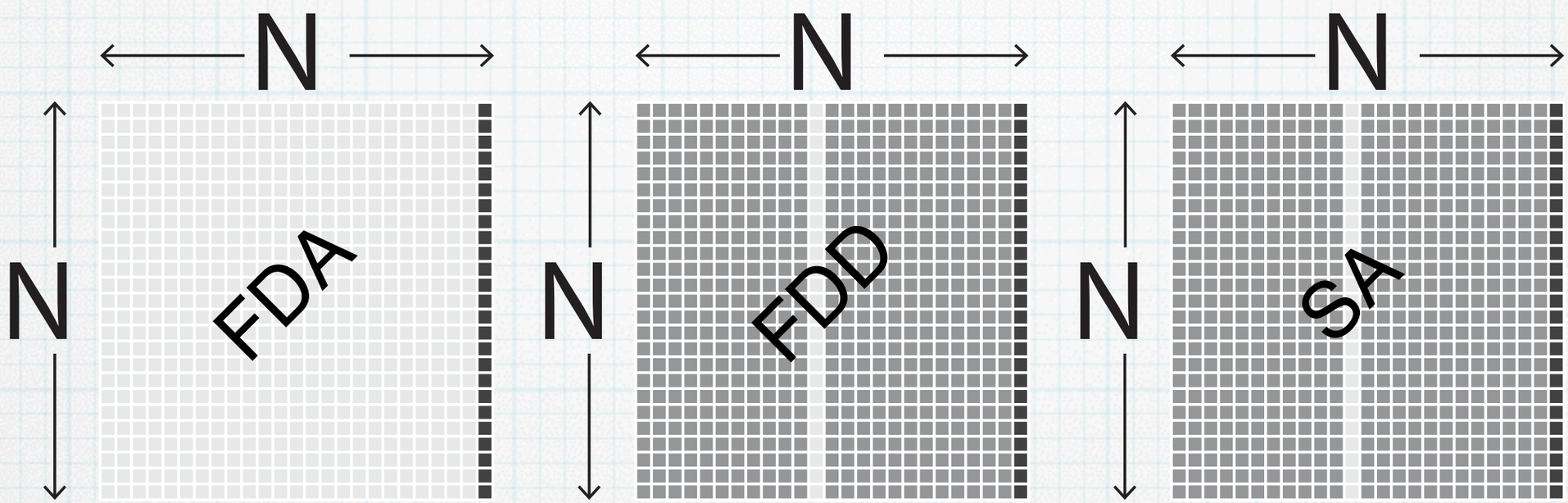
| Per Synapse | |
|---|---|
| Area | 1 |
| Energy | 1+1 |
| Time | 1 |
| A×E×T | 1+1 |

* $E = C_{axo}V_{axo}^2/N + C_{den}V_{den}V_{DD}$
* $T = (C_{axo}V_{axo}/I_{axo})/N$

- Axons are digital but dendrites are analog
- Only one axon may be active at any time

# Summary: Dedicated Synapses

Neuron ■ Synapse ■ Activated synapse

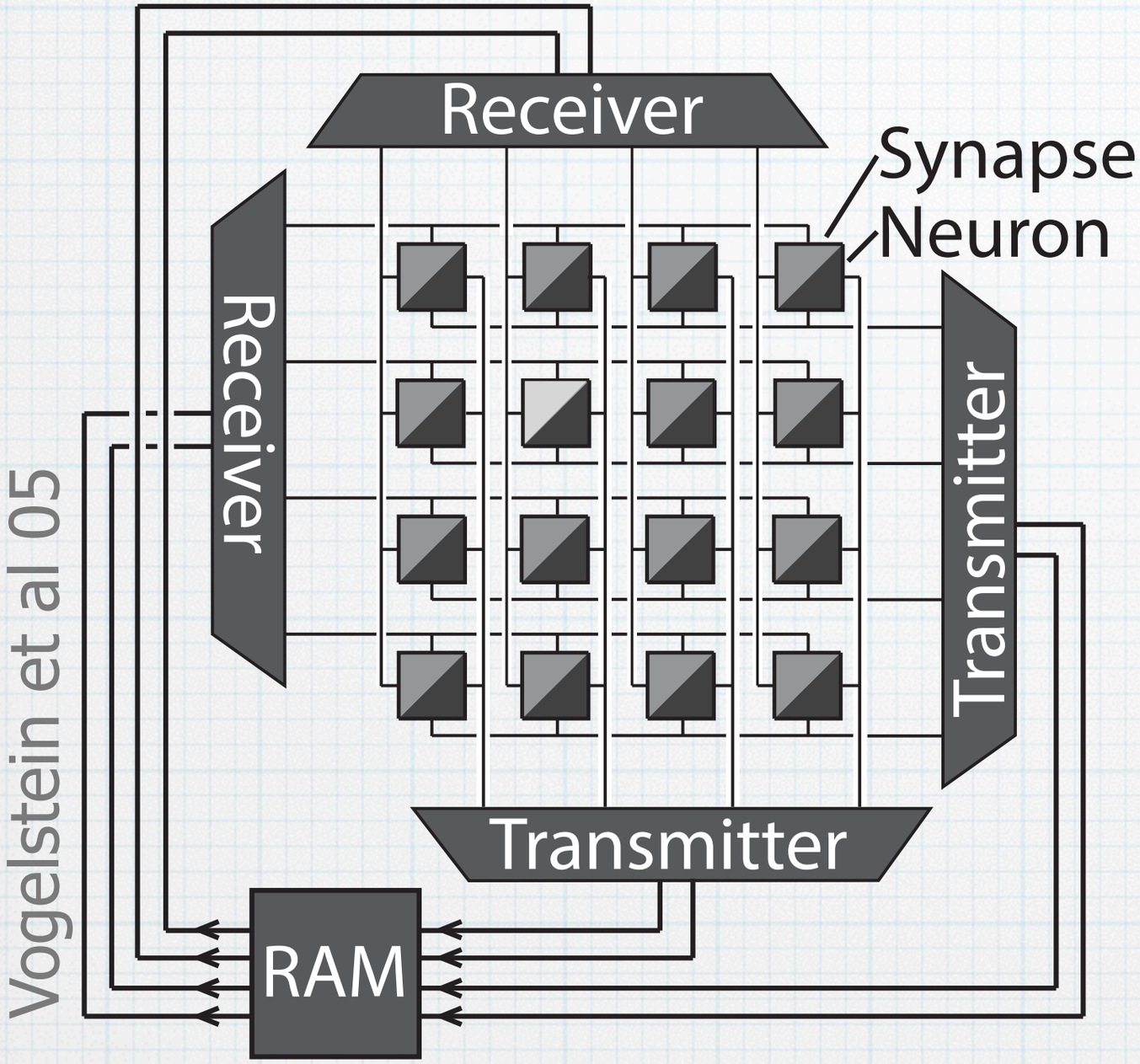



Sivilotti et al 88 Merolla et al 11 Indiveri et al 04

| 1 synapse | FDA | FDD | SA |
|---|---|---|---|
| Area | 1 | 1 | 1 |
| Energy | 1+1 | 1+N | 1+1 |
| Time | 1/N | 1 | 1 |
| A×E×T | 1/N | N | 2 |

AET scales as N, 1, and 1/N for digital, hybrid and analog realizations, respectively.

# Shared Synapse (SS)

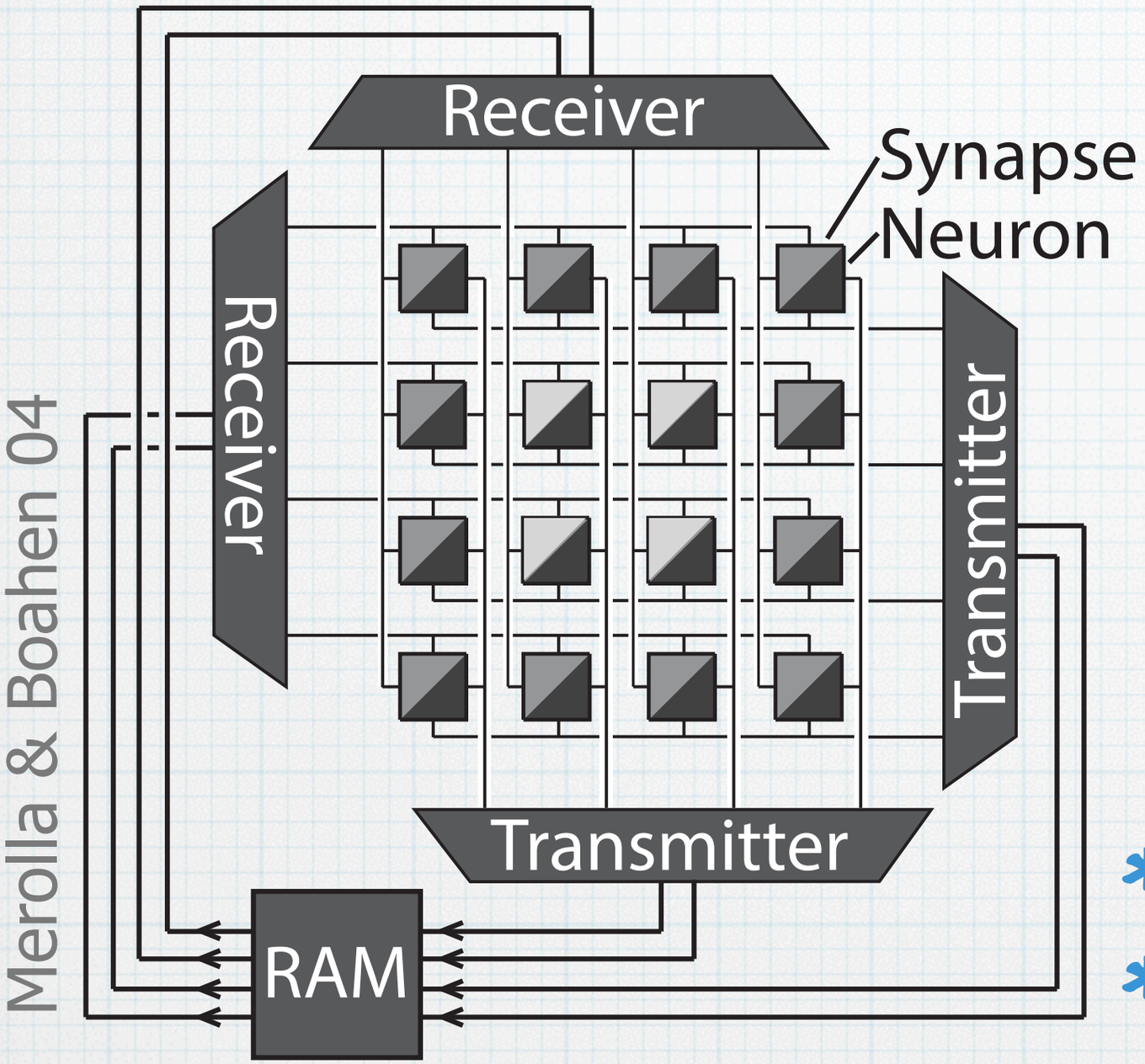



| Per Synapse | |
|---|---|
| Area | 1/N |
| Energy | $2\sqrt{N}$ |
| Time | $\sqrt{N}$ |
| A×E×T | 2 |

* $E = C_{row}V_{DD}^2 + C_{col}V_{DD}^2$
* $T = (C_{row}V_{DD}/I_{axo})$

- A single circuit models all N of a neuron's synapses, so area/synapse = 1/N
- Somas and shared synapses are tiled in a $\sqrt{N}$x$\sqrt{N}$ array, so $C_{row}=C_{col}=\sqrt{N}$

# Shared Dendrite (SD)



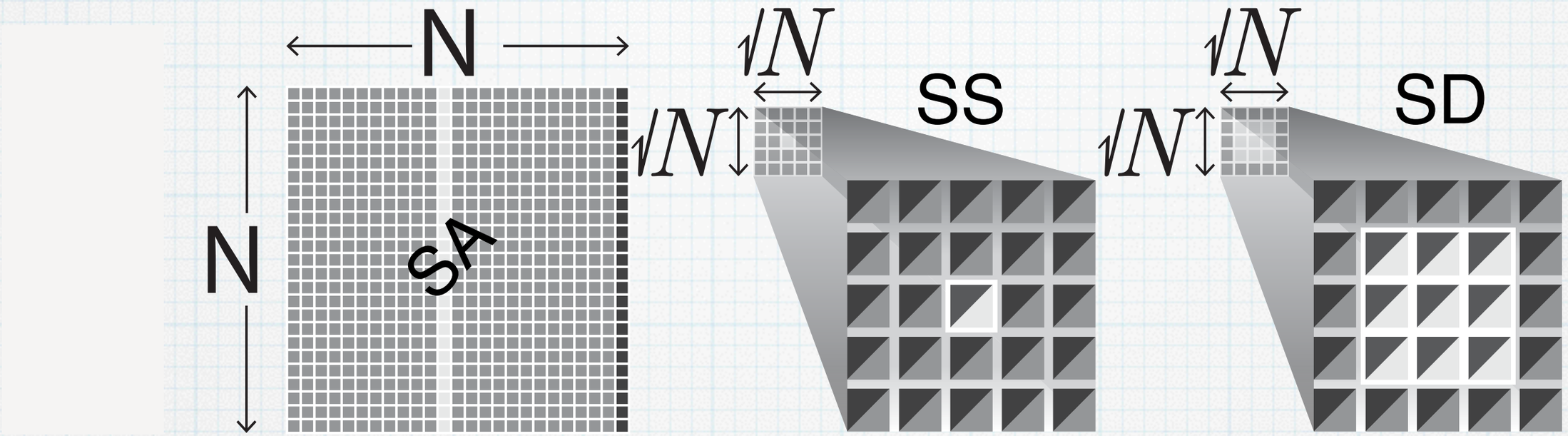

| Per Synapse | |
|---|---|
| Area | 1/N |
| Energy | 2 |
| Time | 1 |
| A×E×T | 2/N |

* $E = (C_{row}V_{DD}^2 + C_{col}V_{DD}^2)/\sqrt{N}$
* $T = (C_{row}V_{DD}/I_{axo})/\sqrt{N}$

- A shared dendrite circuit delivers synaptic input to $\sqrt{N}$ neighboring neurons
- Thus, multiple neurons receive synaptic input (lighter shade) from one spike

# Summary: Shared Synapses

Activated synapse  Neuron+Synapse

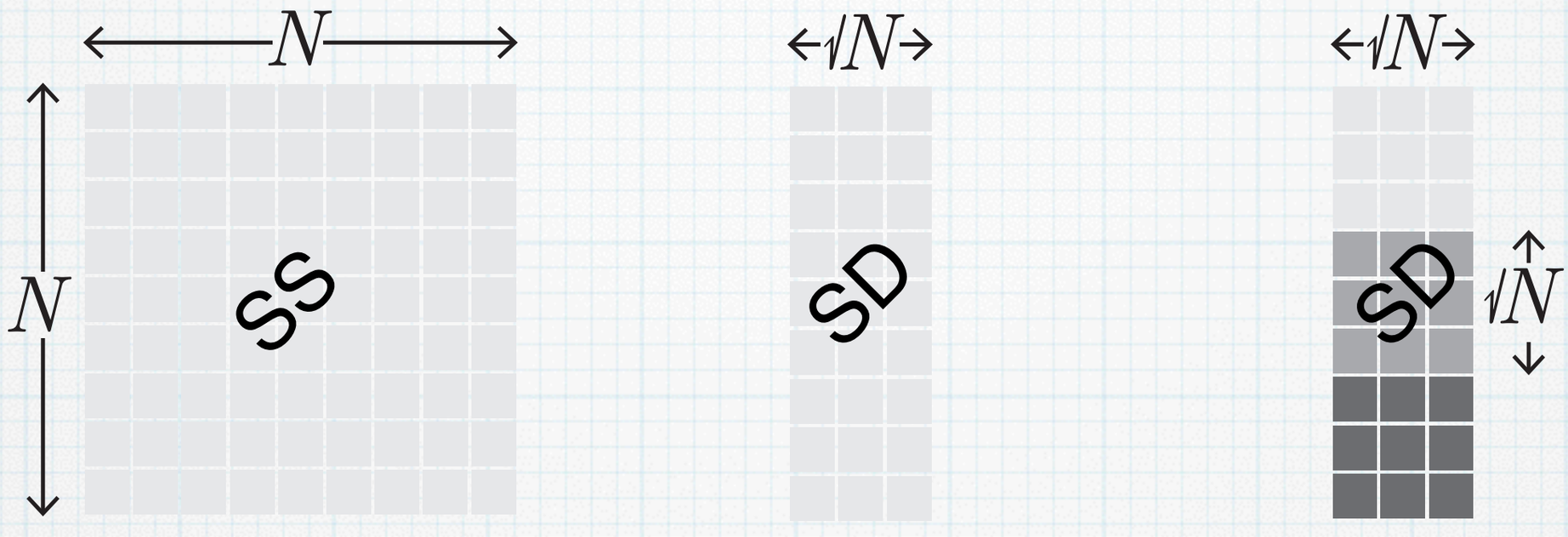



Indiveri et al 04  Vogelstein et al 05  Merolla et al 04

| 1 | SA | SS | SD |
|---|---|---|---|
| Area | 1 | 1/N | 1/N |
| Energy | 1+1 | 2√N | 2 |
| Time | 1 | √N | 1 |
| A×E×T | 2 | 2 | 2/N |

Sharing synapses plus axons doesn't change AET scaling—drops from 2 to 2/N when dendrites are shared as well.

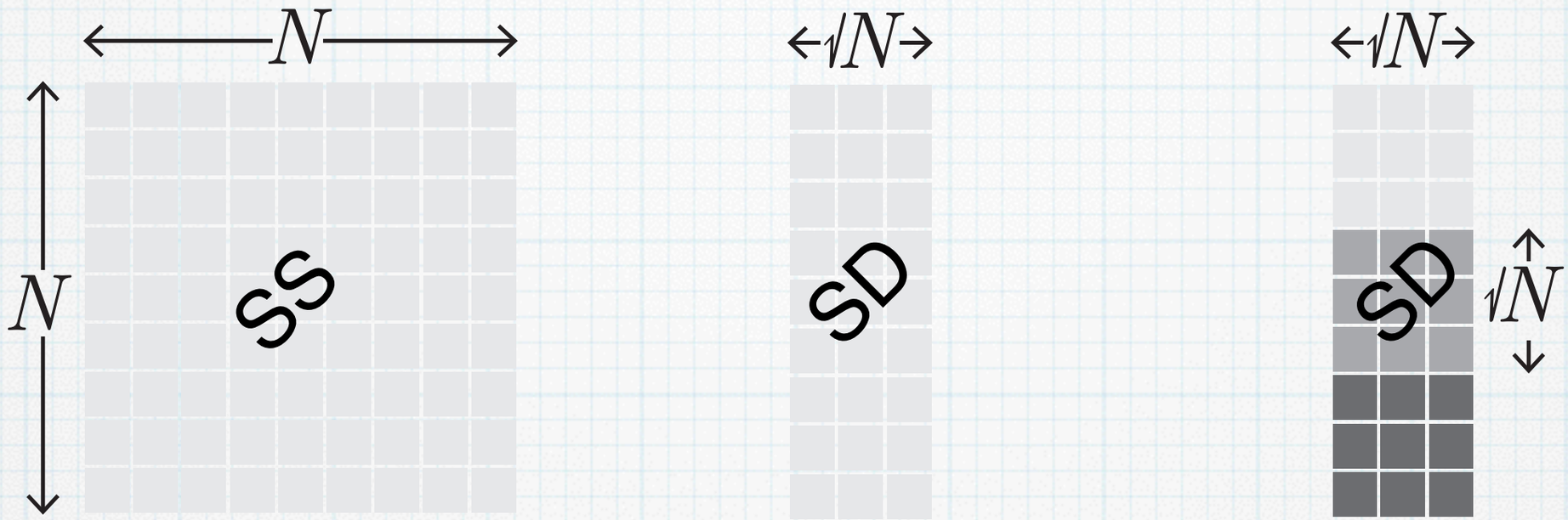
# SS & SD's RAM Costs




| 1 synapse | Shared Synapse | Shared Dendrite | Banked |
|---|---|---|---|
| Area | 1 | √N/N | √N/N |
| Energy | 1+N | (√N+N√N)/N | (√N+√N√N)/N |
| Time | 1+1 | (√N+N)/N | (√N+√N)/N |
| A×E×T | 2N | 1 | 1/N |

- Assumes throughput is not limited by I/O (i.e., RAM is on-chip)
- Small banks makes AET cost comparable to Shared Dendrite

# Summary: RAM Costs



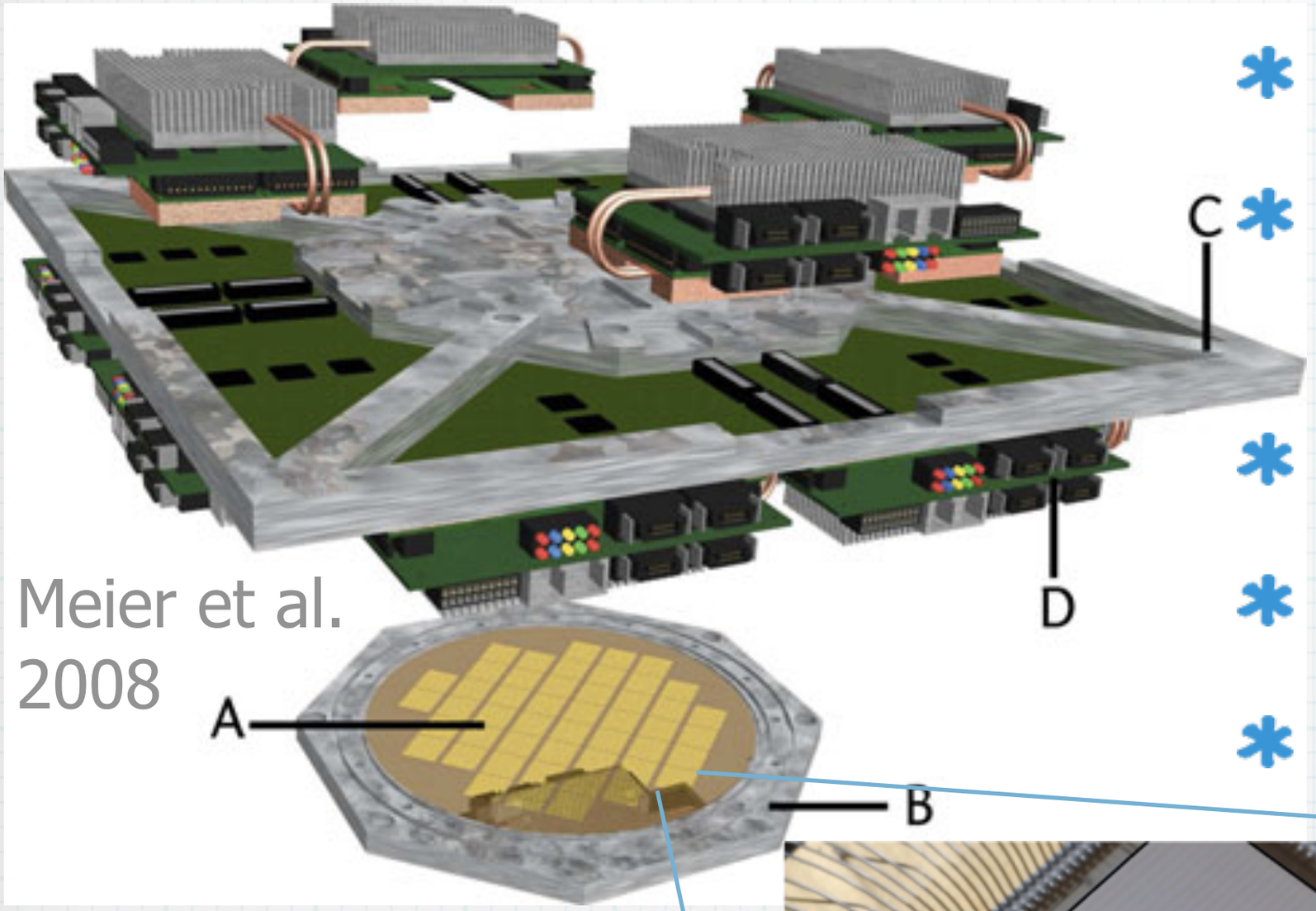

| 1 | SS | SD | Bank |
|---|---|---|---|
| Area | 1 | 1/√N | 1/√N |
| Energy | 1+N | 1/√N+√N | 1/√N +1 |
| Time | 1+1 | 1/√N +1 | 2/√N |
| A×E×T | N | 1 | 2/N |

SS' RAM AET cost scales as N, compared to 1 for its neuron array.

SD's RAM AET scales as 1, compared to 2/N for its neuron array.

* When Shared Dendrite is paired with Heavily Banked RAM, its AET cost (2/N) is as low as Fully Dedicated Analog's

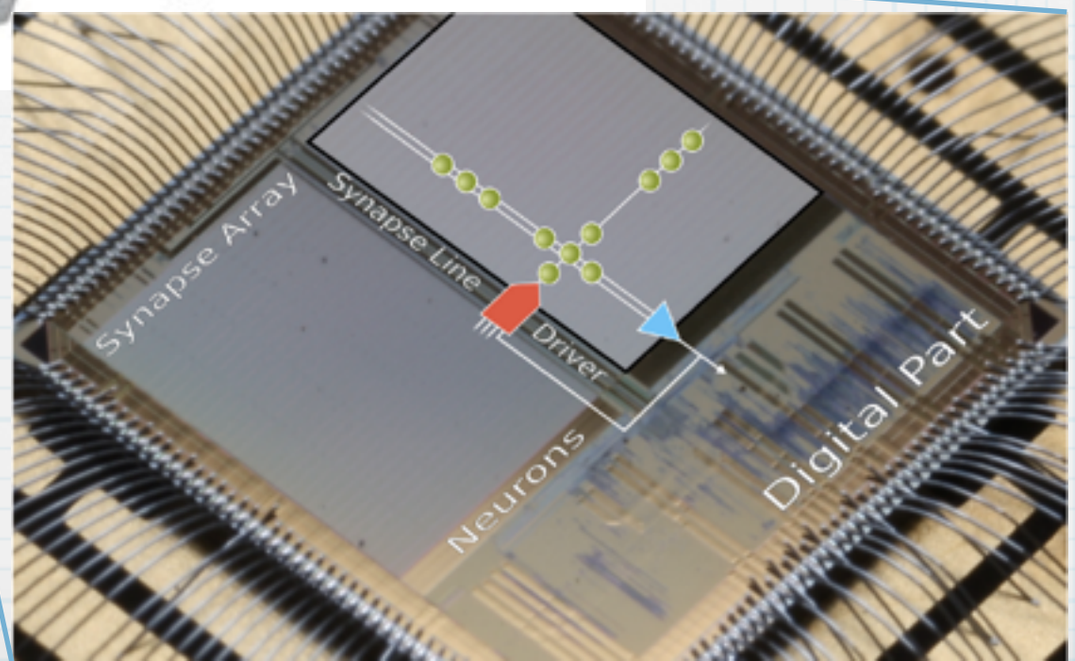
# HiCANN/FACETS/BrainScales




Meier et al. 2008

* Fully Dedicated
* Analog synapses, dendrites, and soma
* Digital axons
* Plastic synapses
* 1000x real-time

Wafer:
384 Chips
200K neurons
45M synapses
$10^3$ x real-time
800W (est.)

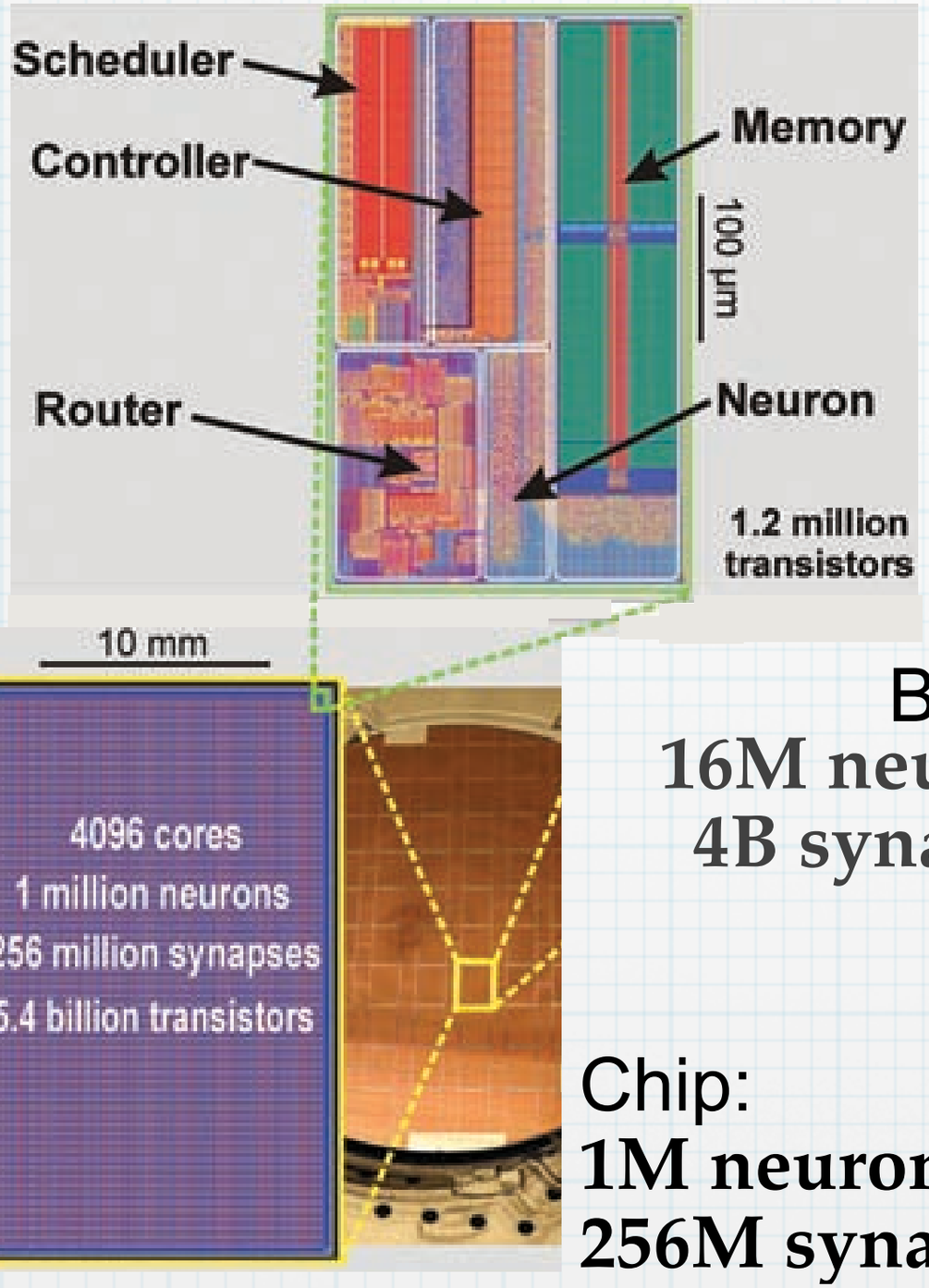



Chip:
384 neurons
256 synapse/ea
$10^3$ x real-time
3W (est.)

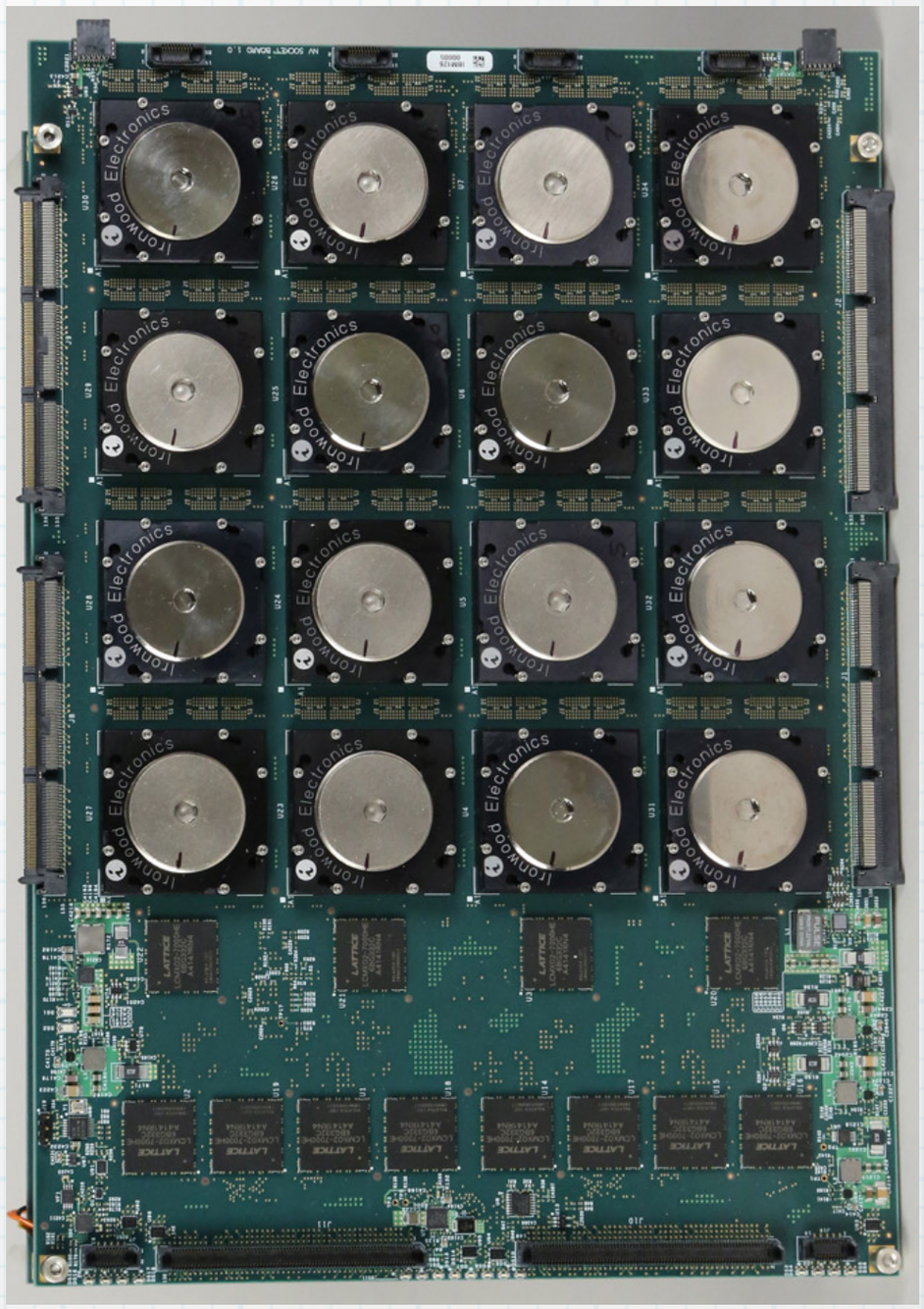
Meier et al 12

# IBM TrueNorth

Core:
256
neurons
64K
synapses




Board:
16M neurons
4B synapses
7.2W

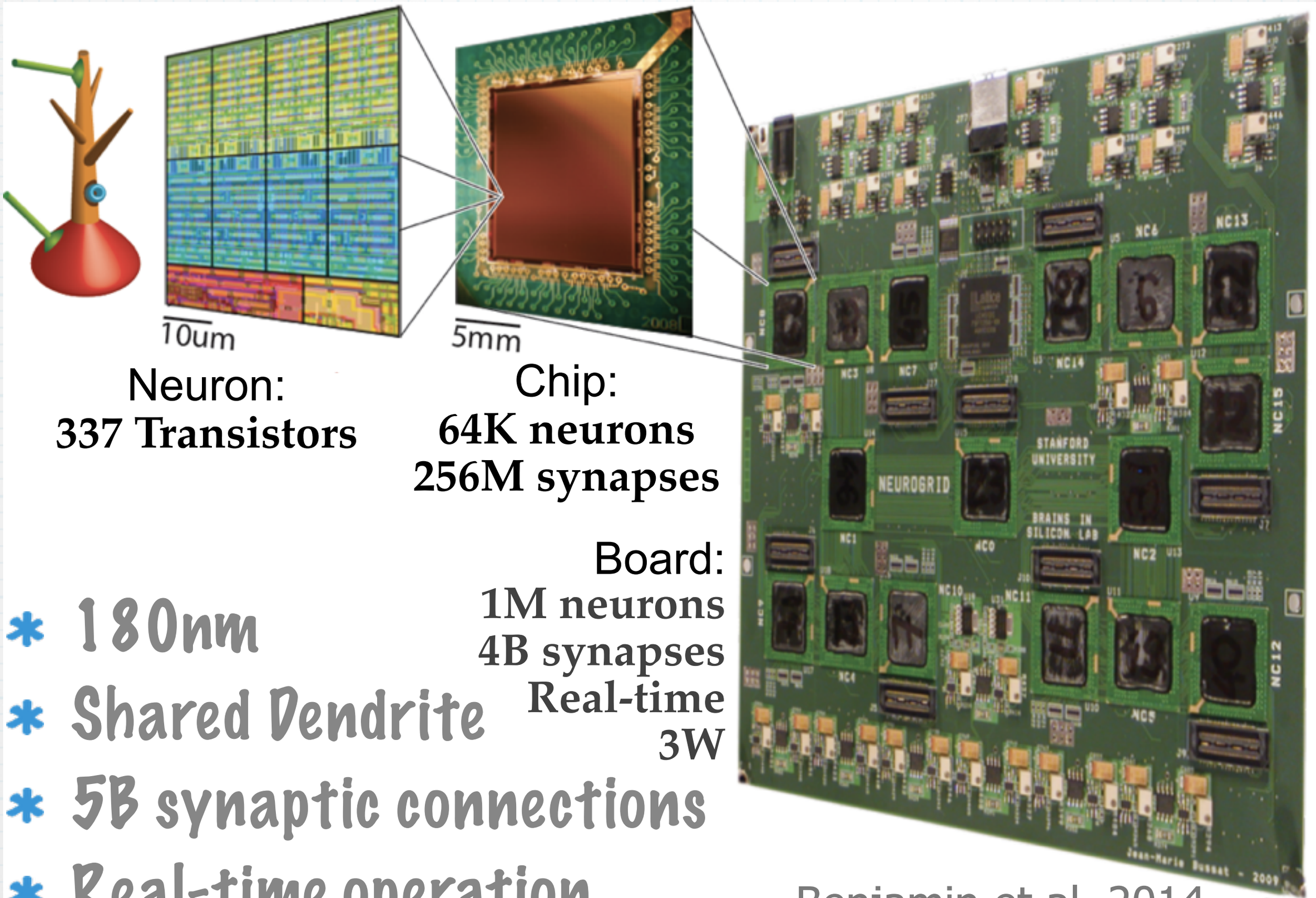


Chip:
1M neurons
256M synapses

* Fully Shared
* All digital
* 28nm
* Real-time operation

Merolla et al 14

# Neurogrid




Neuron:
337 Transistors

Chip:
64K neurons
256M synapses

Board:
1M neurons
4B synapses
Real-time
3W

* 180nm
* Shared Dendrite
* 5B synaptic connections
* Real-time operation

Benjamin et al. 2014

# Chip-Level Comparison

| | $A$ ($\mu$m$^2$) | $E$ (pJ)$^\dagger$ | $T$ (ps) | $A{\cdot}E{\cdot}T$ | $S$ | Synapse Spec |
|---|---|---|---|---|---|---|
| HICANN | 436 | 198 | 2.9 | 250K | 224 | 4-bit plastic |
| GoldenGate | 256 | 4.0K | 29.5 | 30.6M | 1024 | 1-bit dedicated |
| Neurogrid | 0.63 | 119 | 62.5 | 4.69K | 4096 | 13-bit shared |

$^\dagger$ $E$ includes dynamic and static power (for a 10 spike/s/neuron simulation).

For comparison with HICANN and Neurogrid, GoldenGate's A=16um$^2$, E=1.9nJ, and T=3.9pS where scaled from a 0.85V–45nm process to a 1.8V–180nm process using general scaling laws [Rabaey&Chandrakasan02].

Benjamin et al., 2014

# Summary

* Dedicated synapses:
  AET scales as N, 1, and 1/N for digital, hybrid and analog realizations, respectively (per synapse).

* Shared synapses:
  AET scales as 1—same as hybrid.
  RAM scales as N—same as digital.

* Shared dendrites:
  AET scales as 1/N—same as analog.
  RAM scales as 1—same as hybrid.
  Banking cuts this to 1/N—same as analog.