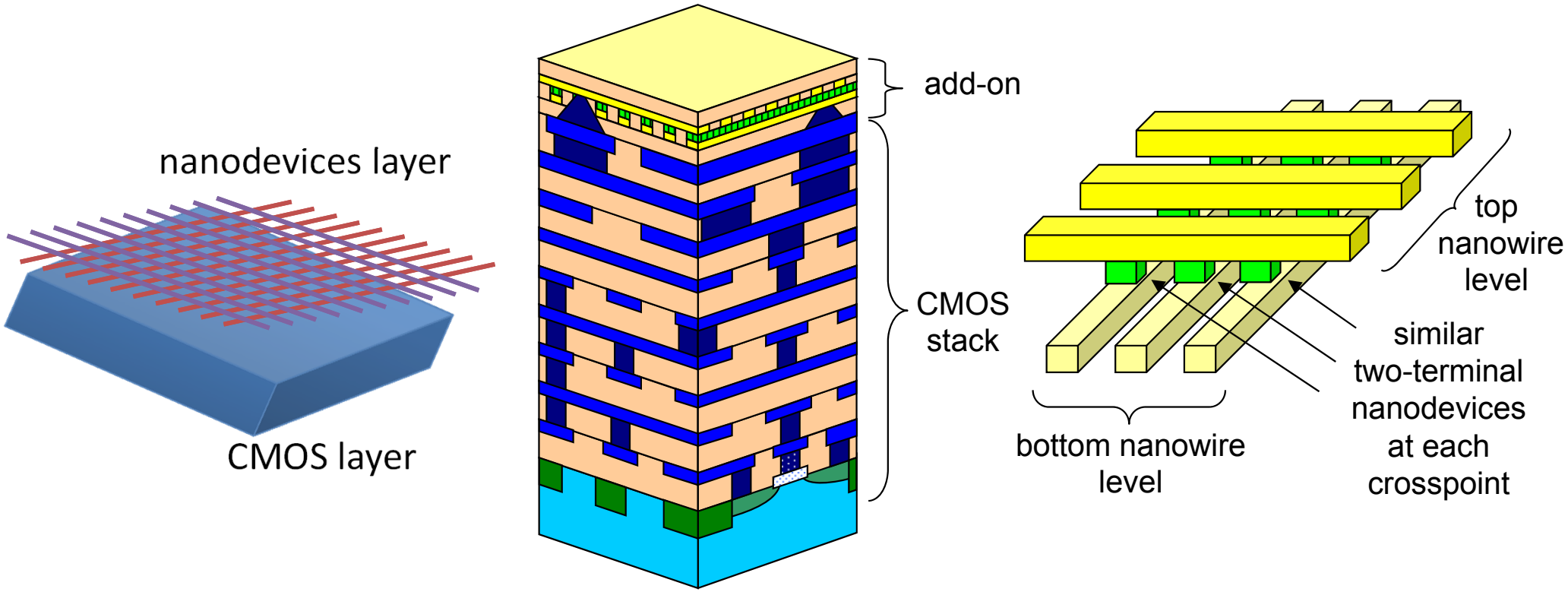
# Integration, Architecture, and Applications of 3D CMOS-Memristor Circuits

K.-T. Tim Cheng and Dimitri Strukov

Univ. of California, Santa Barbara

**ISPD 2012**

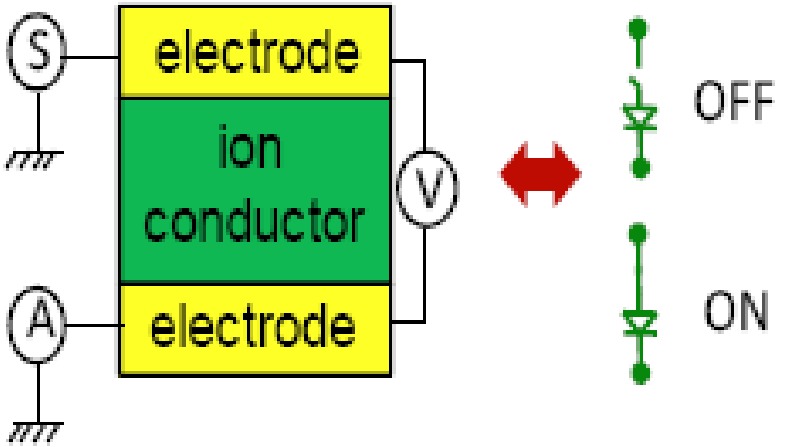
# 3D Hybrid - CMOS/NANO

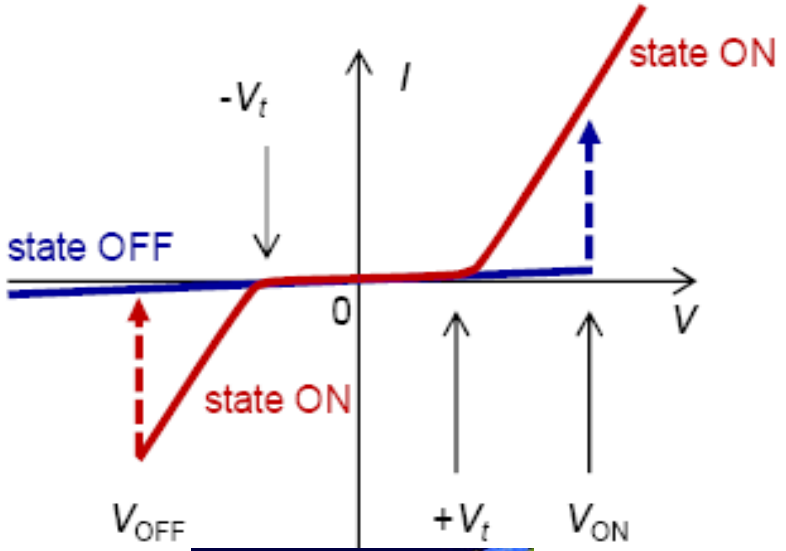



- CMOS stack + nano add-on
- nanowire crossbar of two-terminal devices (memristors)

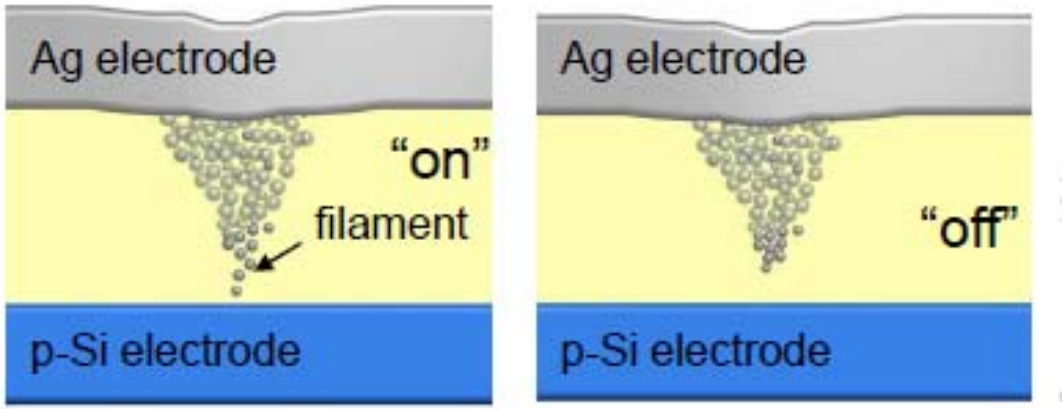
# Resistive Switching "Memristive" Devices

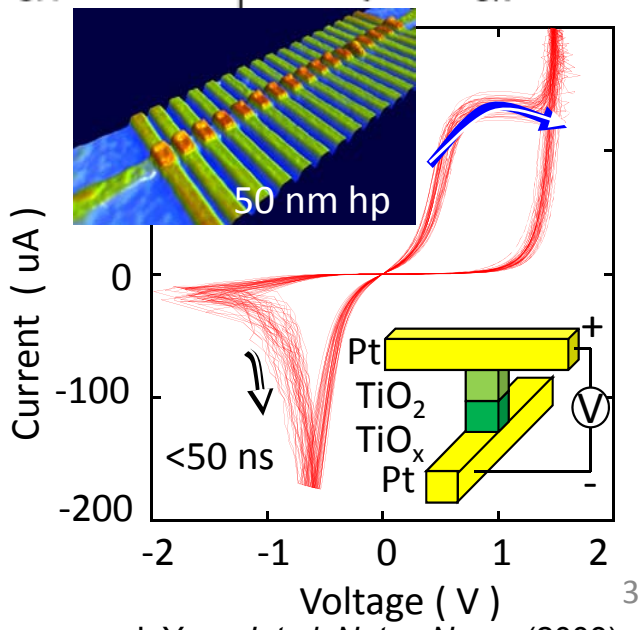
(*latching switches*, a.k.a. *resistive switches*, a.k.a. *programmable diodes, a.k.a. memristive switches*)



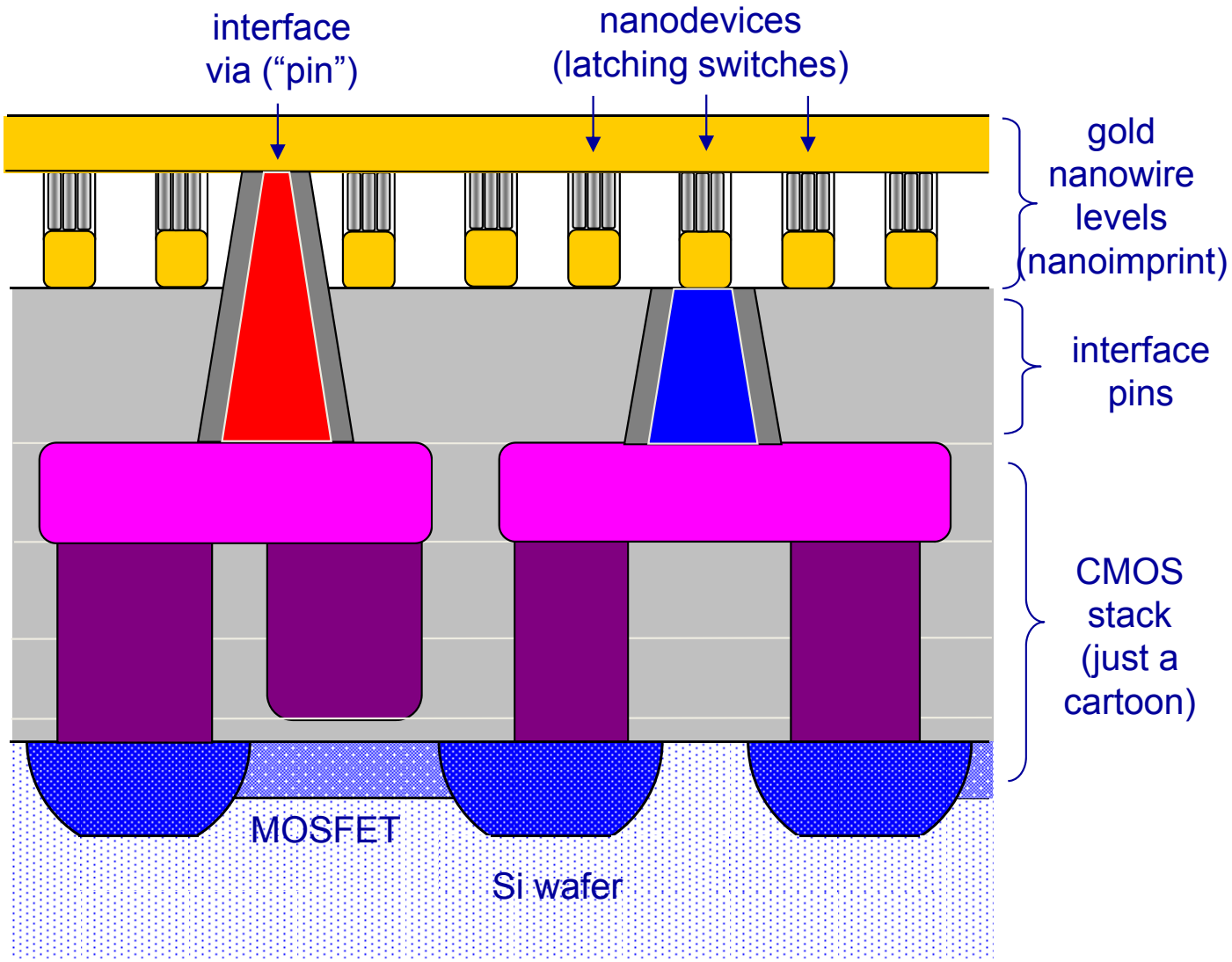

**+ Wide range of material systems and physical phenomena**



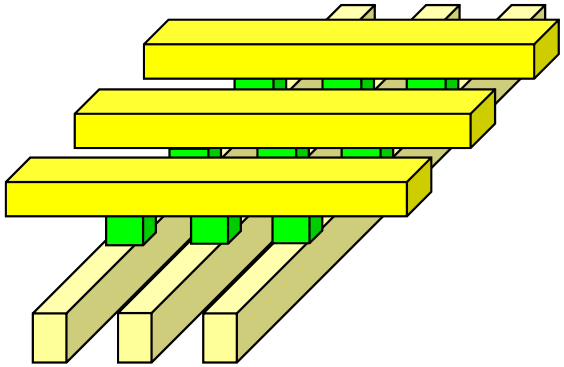




J. Yang *let al. Natue Nano,* (2008)

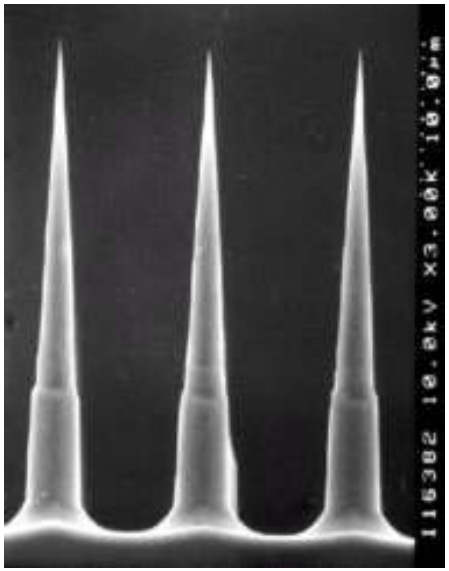
# Area-Distributed "CMOL" Interfaces




K. Likharev (2004, 2005); D. Strukov and K. Likharev (2006)

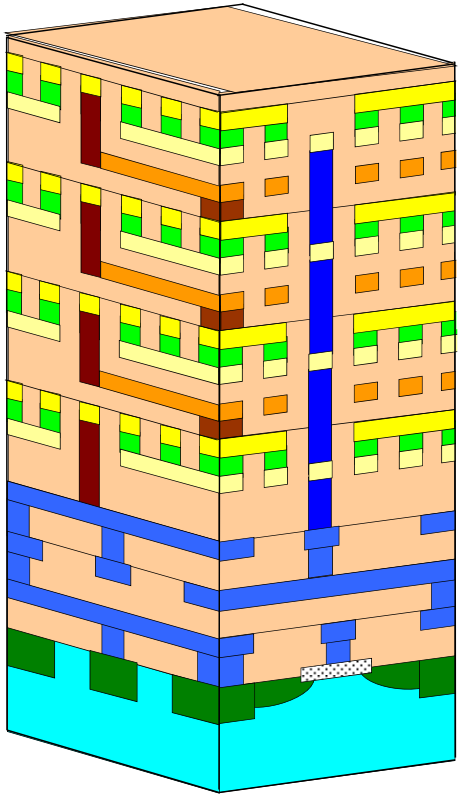


Tip radii 2-10 nm



http://www.oxfordplasma.de/process/sibo_wtc.htm

# AFOSR-MURI
# HyNano: 3D Hybrid CMOS-Nano Circuits



## Overall Goal

**Develop and demonstrate applications, architectures, circuits, and device technologies for 3D hybrid circuits which integrate CMOS subsystems with nano-memristors**

# The HyNano Team

Michael Chabinyc
Materials, UCSB

Tim Cheng
ECE, UCSB
(Director)

Marivi Fernandez-Serra
Physics, Stony Brook

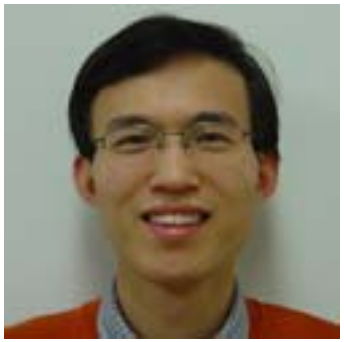
Konstantin K. Likharev
Physics, Stony Brook

Wei Lu
EECS, Michigan

Susanne Stemmer
Materials, UCSB

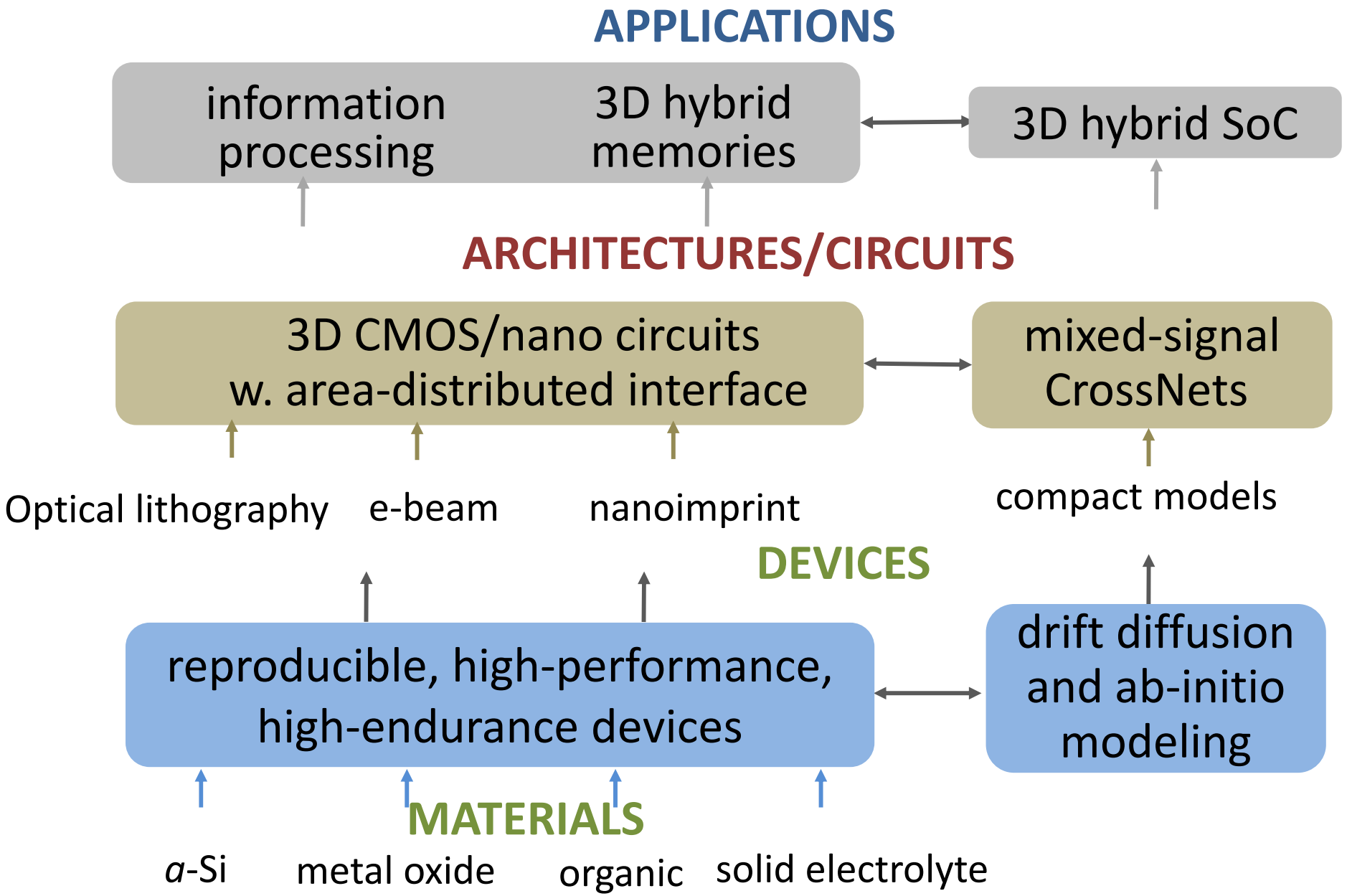
Dmitri Strukov
ECE, UCSB

Luke Theogarajan
ECE, UCSB



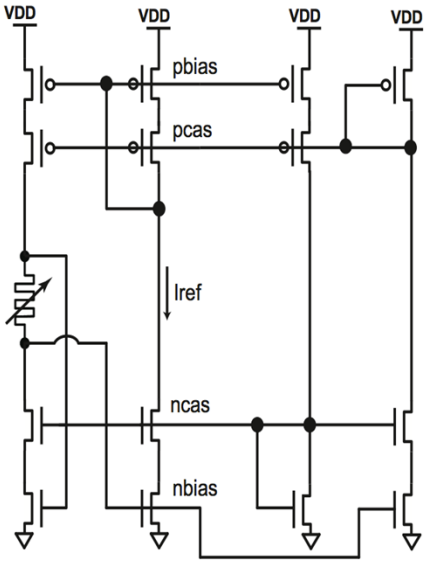
Qiangfei Xia
ECE, UM Amherst

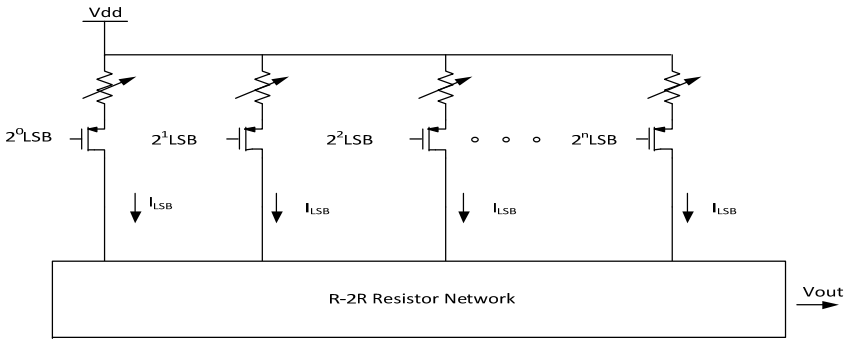
# Project Overview

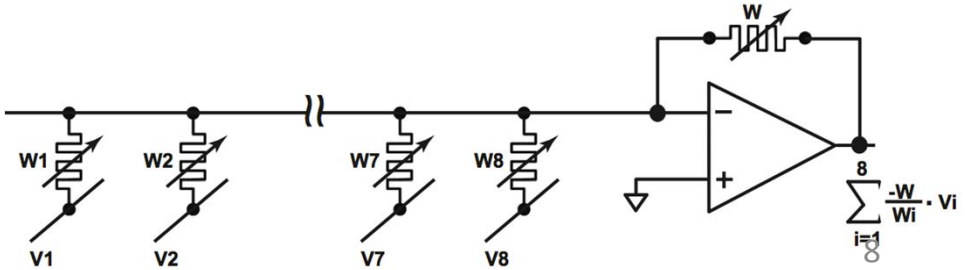


# Thrust Area #1: Application/Architecture/Ckt Exploration

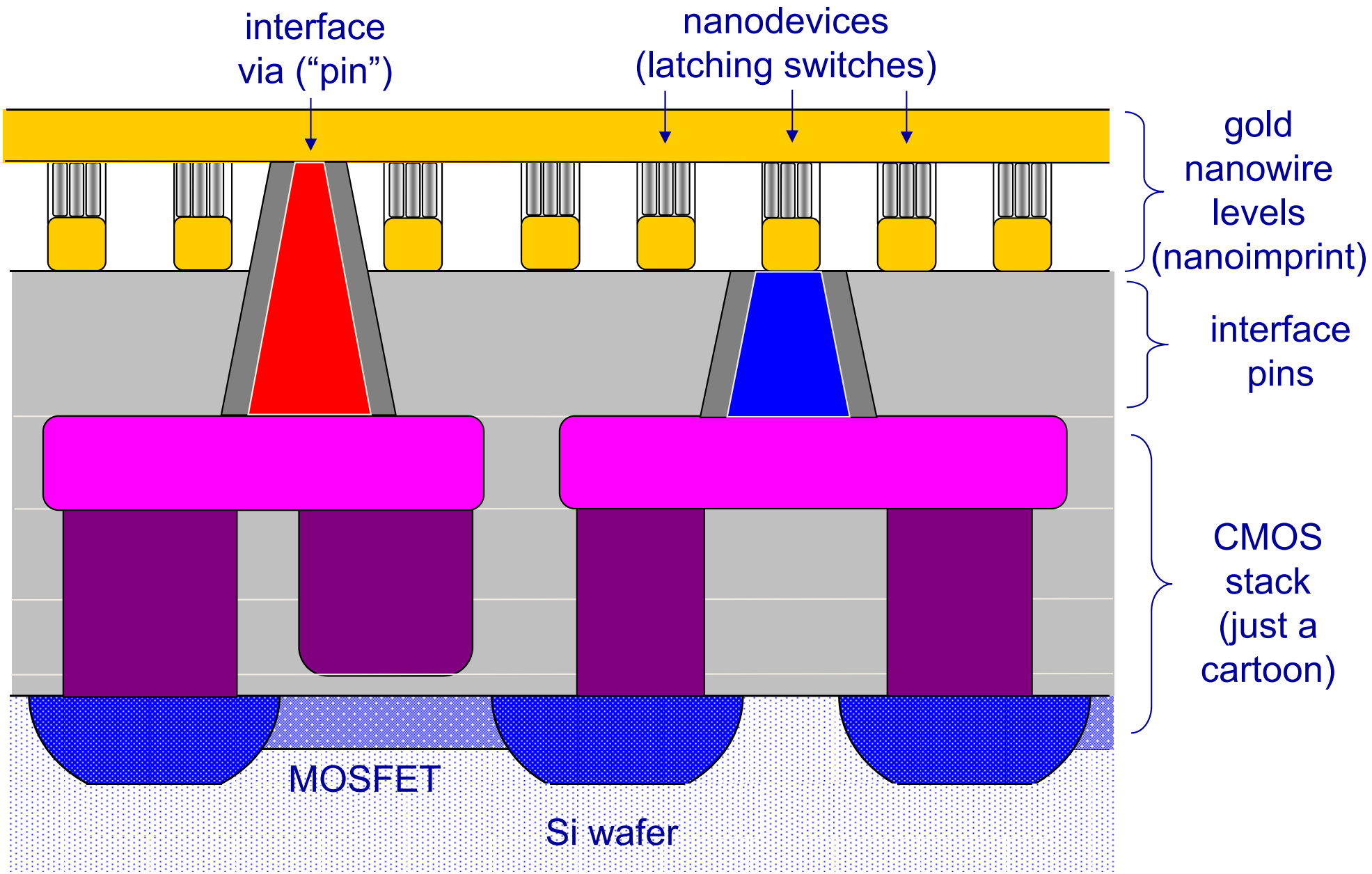
- **Memory arrays for high-performance computing**
- CMOL-based FPGA
- Neuromorphic networks for bio-inspired information processing
- Evolvable analog circuits
- Tunable bias network for analog design
- Weighted multiply and add circuits
- High precision Digital-to-Analog converter



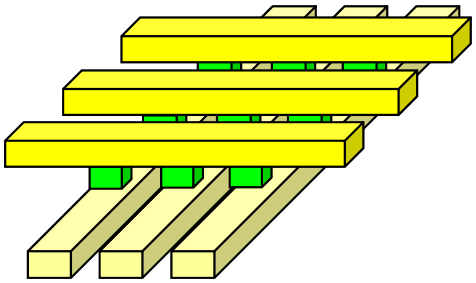



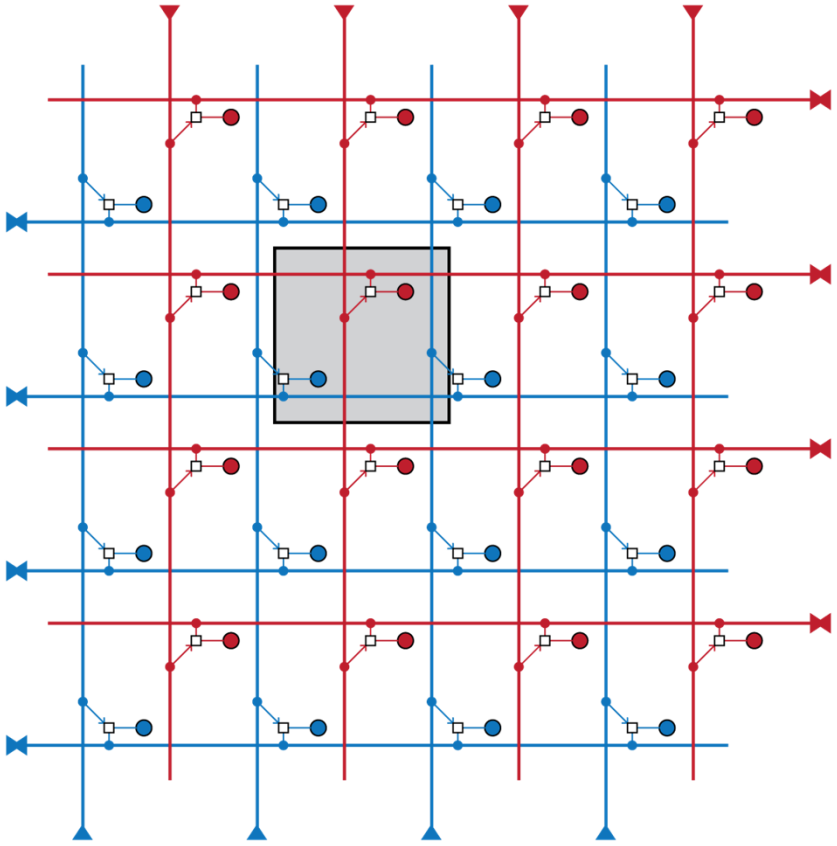
# “CMOL” Interface – Integrating CMOS with Crossbar Memory Array



# Addressing Crossbar Memory Array

- There are two types of pins – Blues and Reds
- Each array of pins has its own decoding scheme



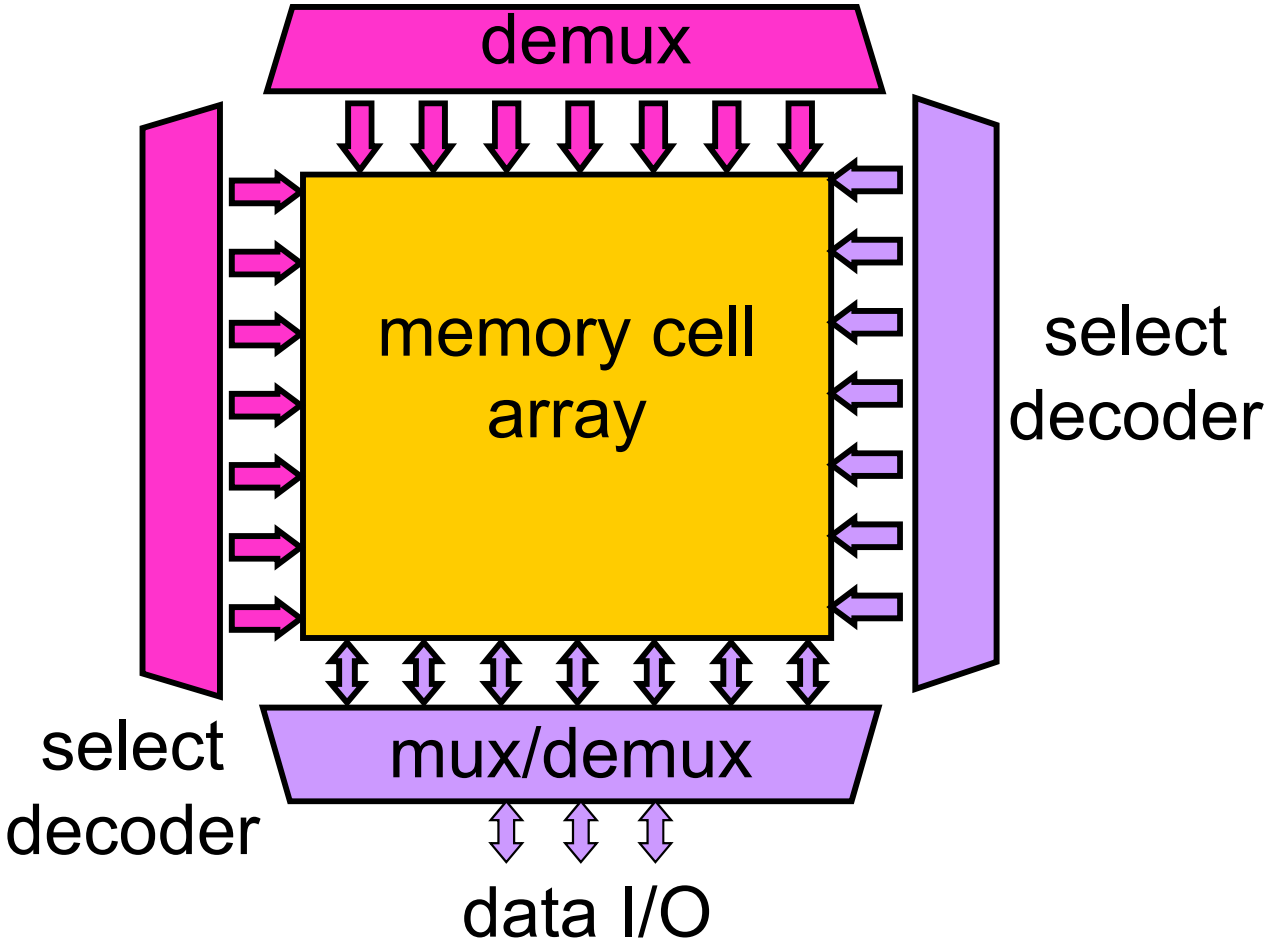


Double decoding scheme:

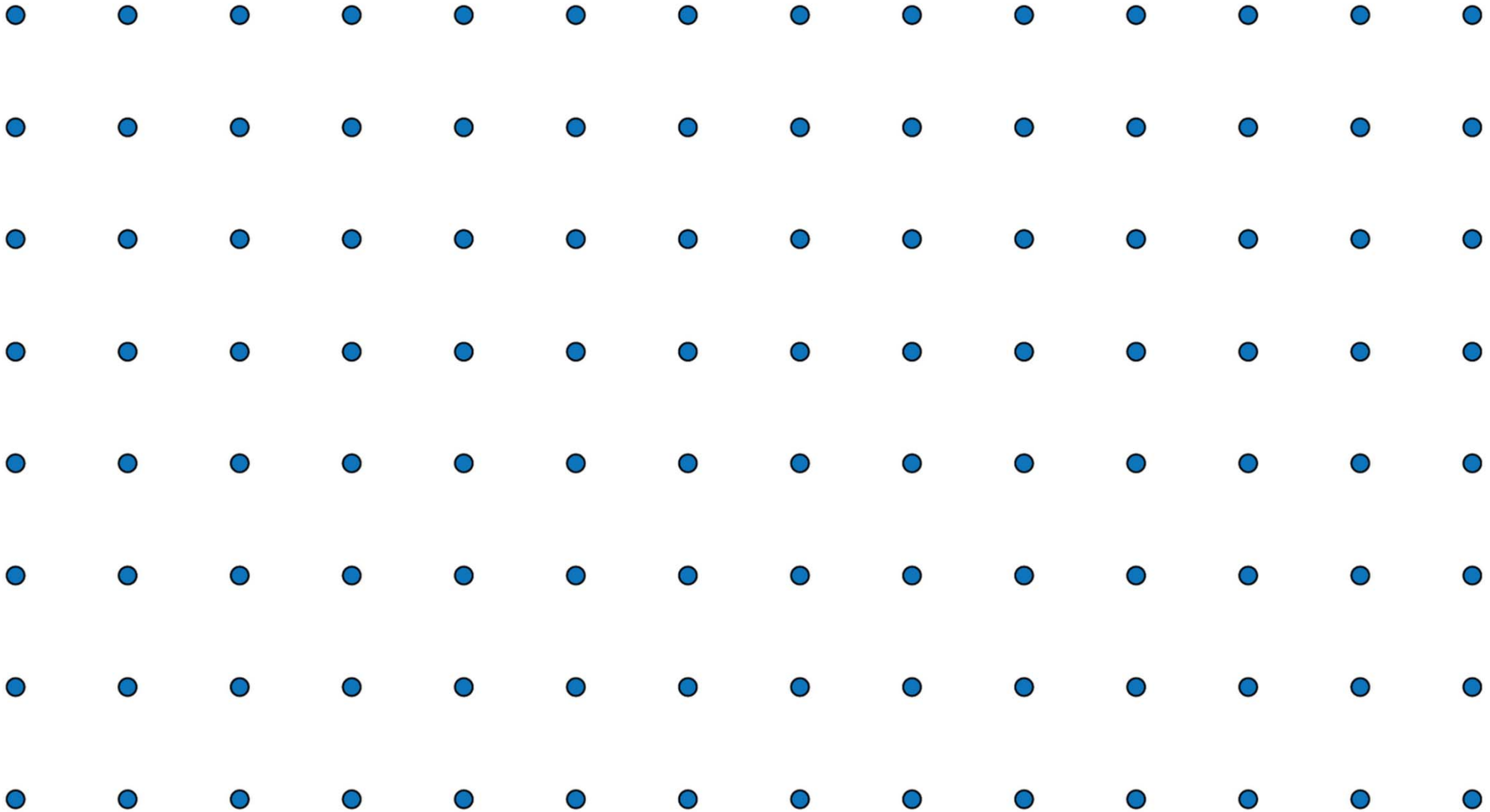
- An array of $N^2$ blue pins uniquely accessed with 2N control signals.
- Another 2N control signals for the corresponding $N^2$ red pins

# Double Decoding Scheme

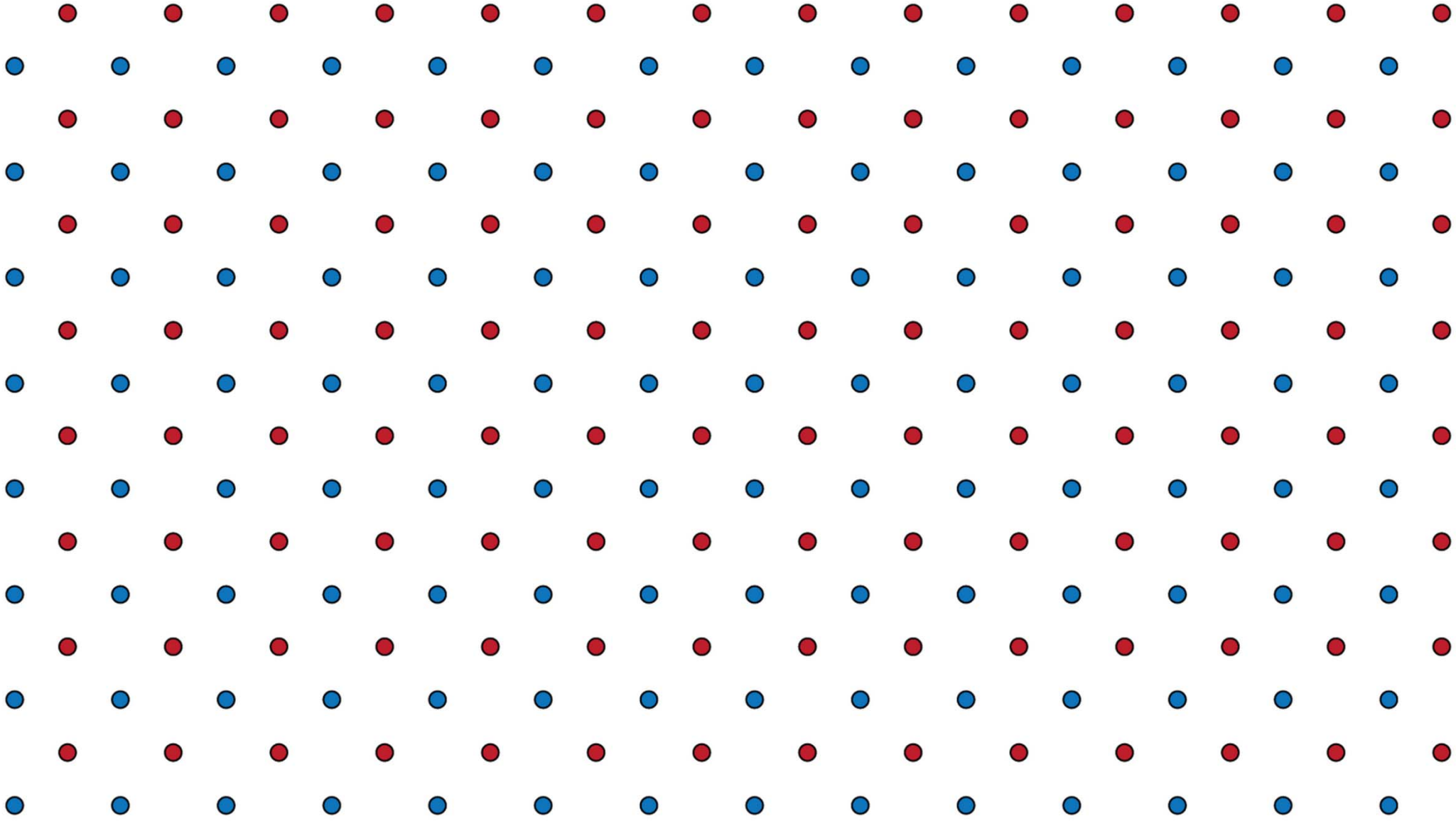
- Four decoders:



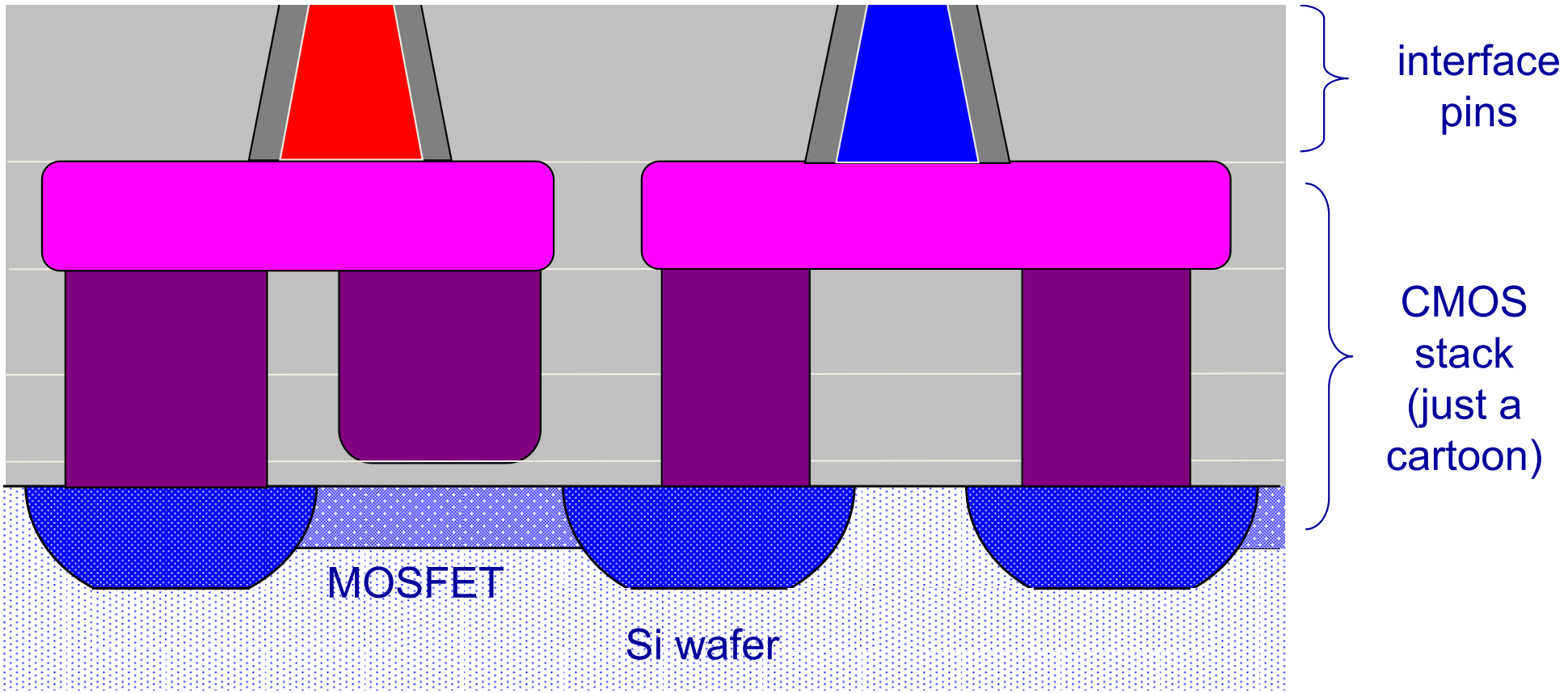
# Crossbar Construction – Top View



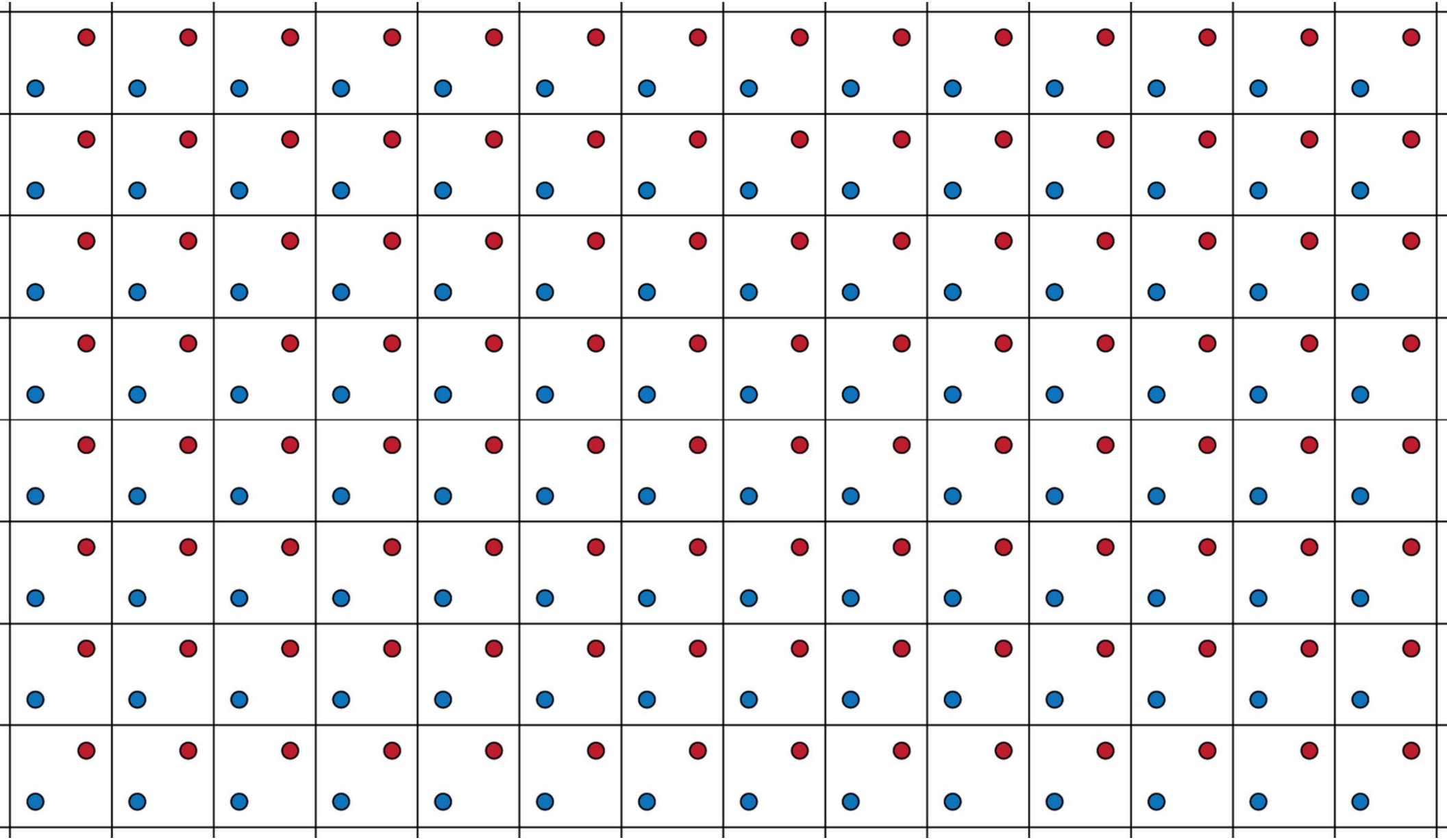
# Crossbar Construction – Top View



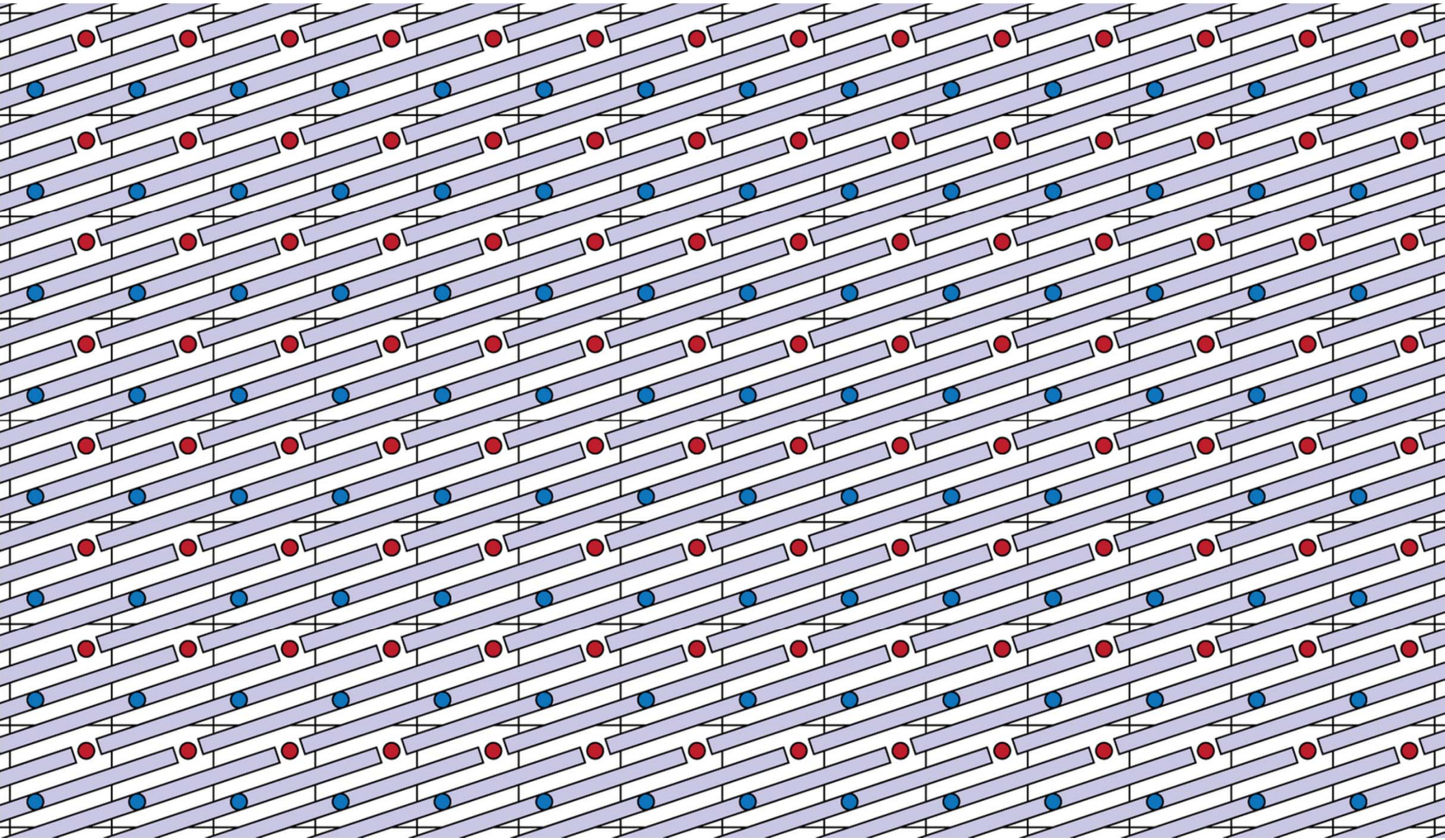
# Crossbar Construction – Side View



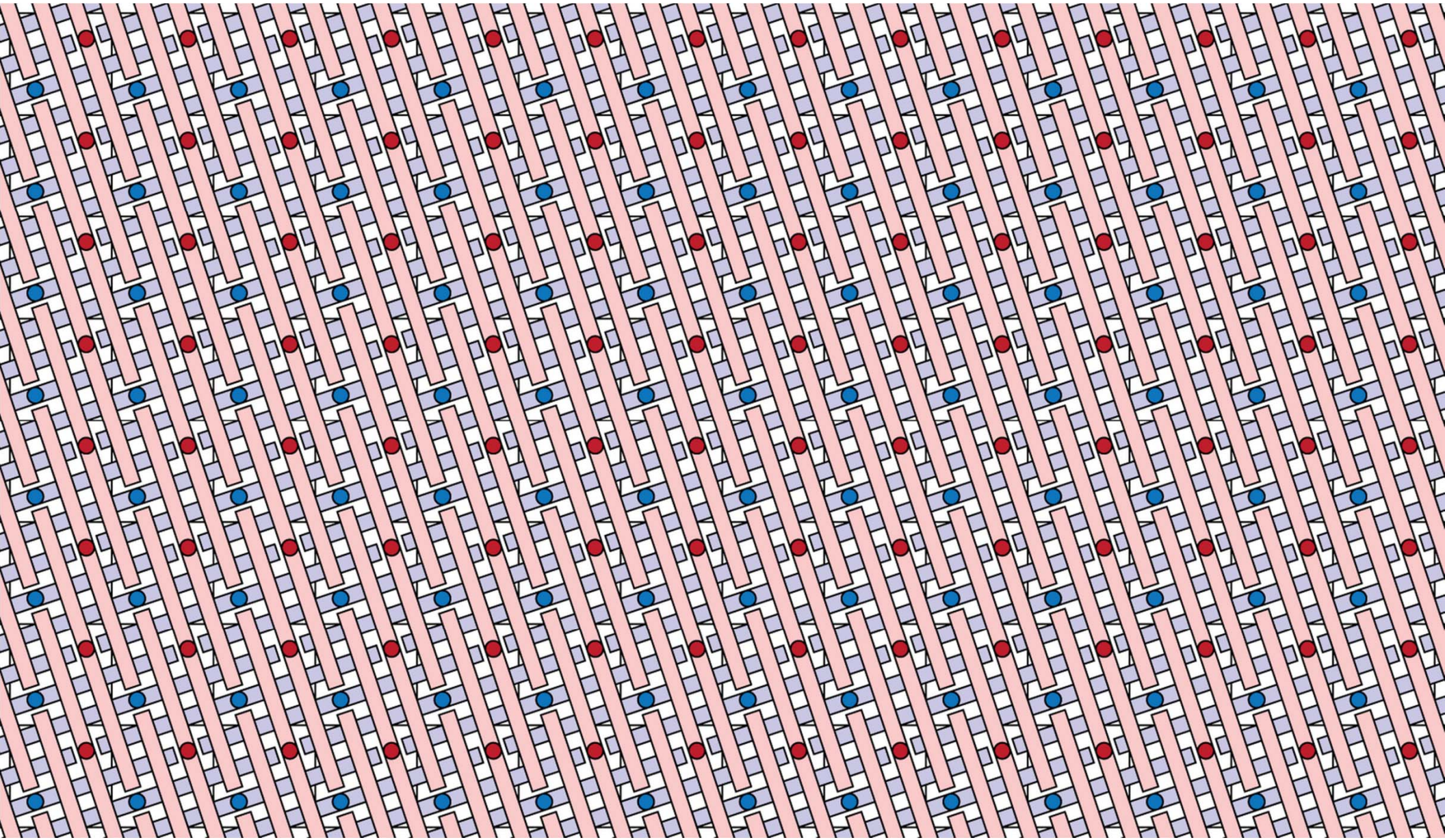
# Crossbar Construction – Top View



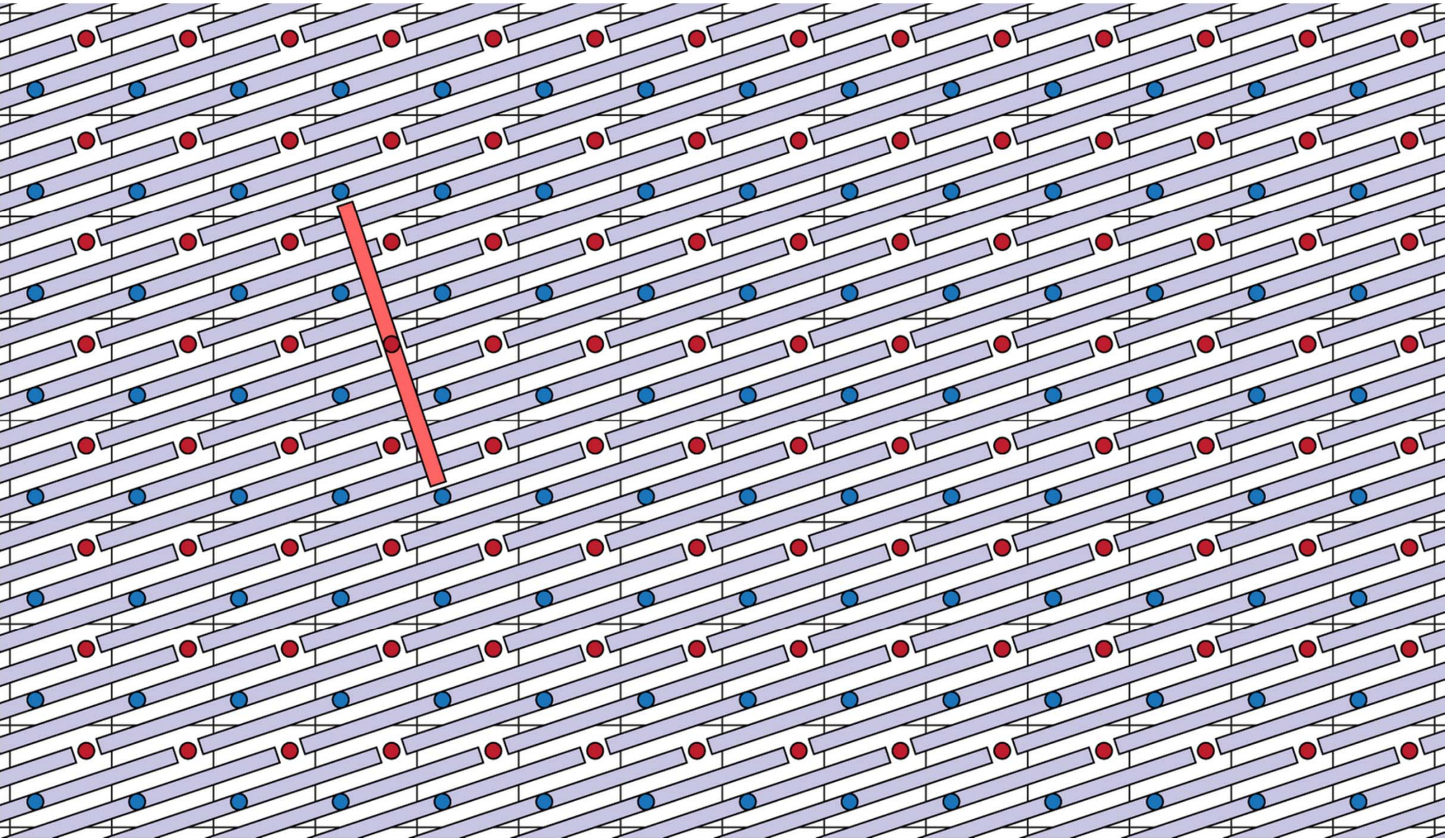
# Crossbar Construction – Bottom Level



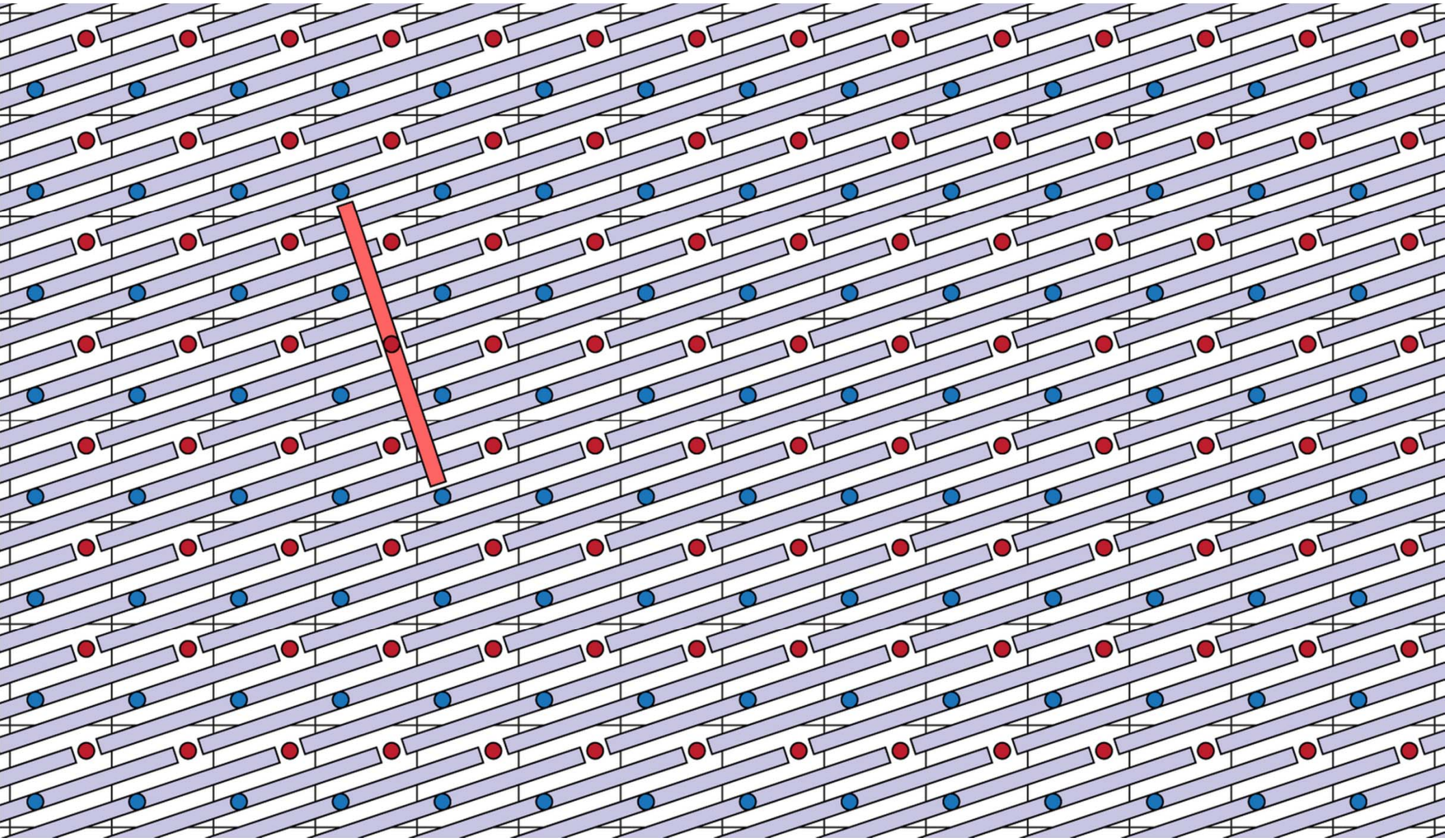
# Crossbar Construction – Top Level



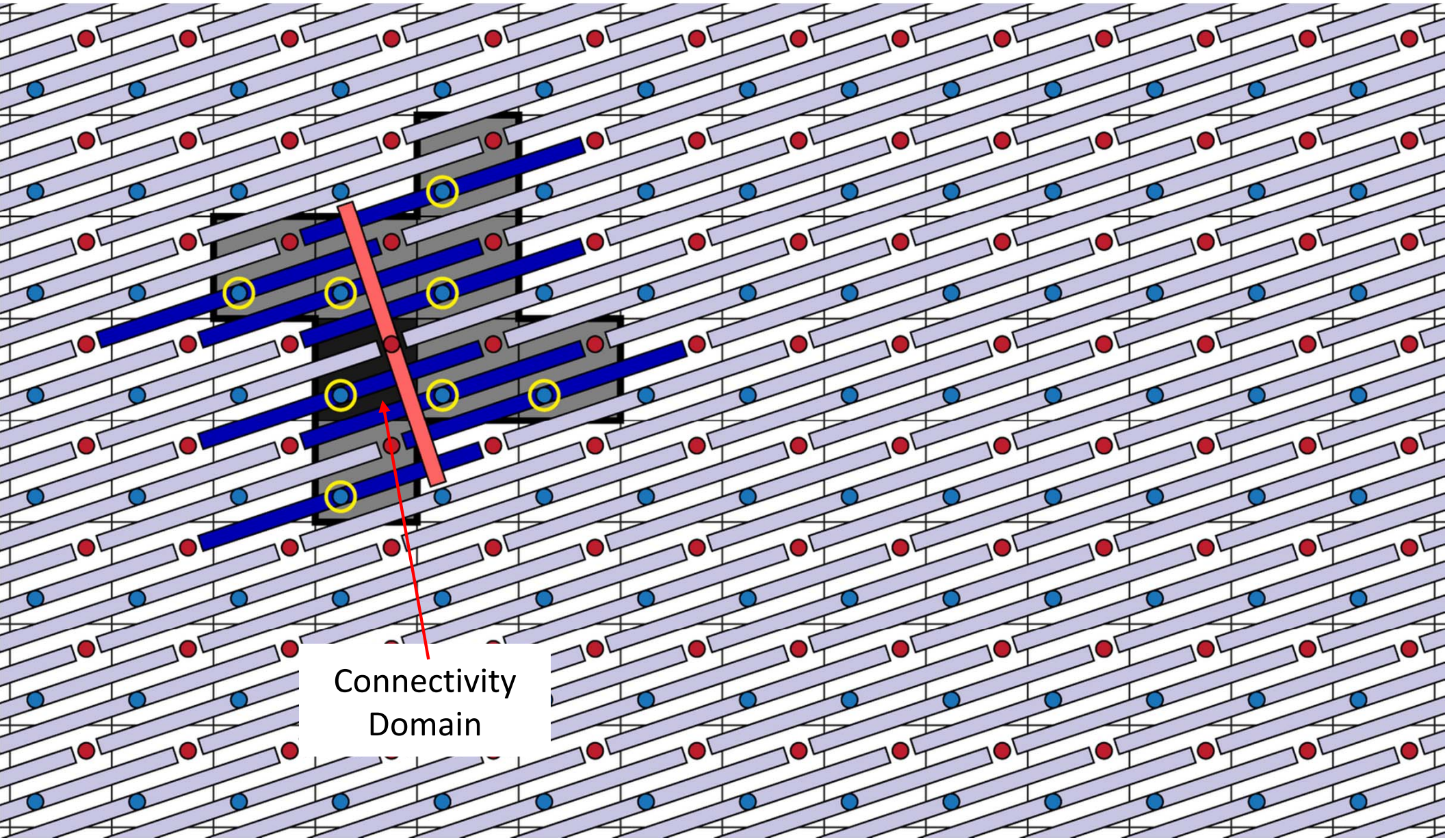
# Crossbar Construction



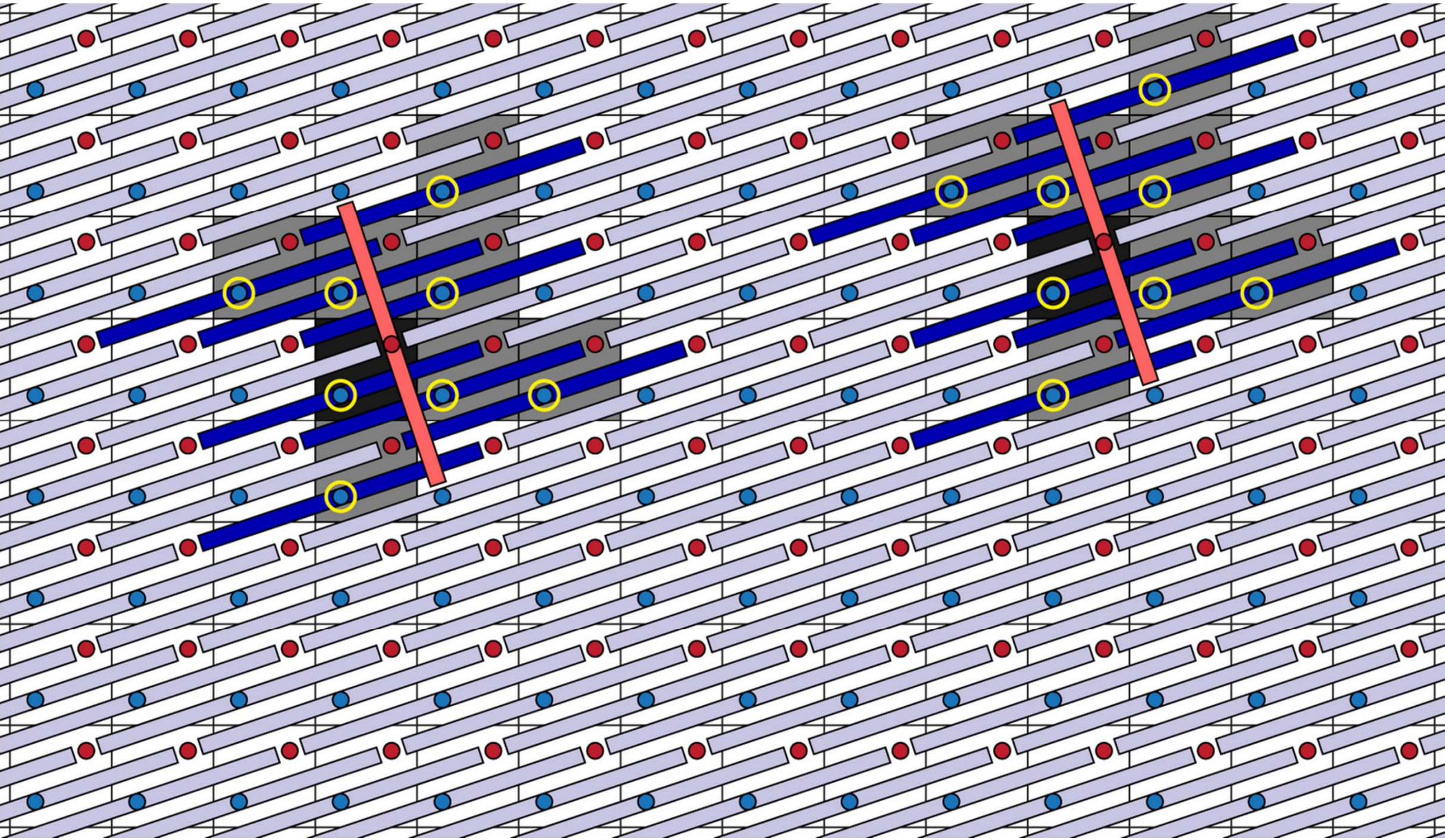
# Crossbar Construction



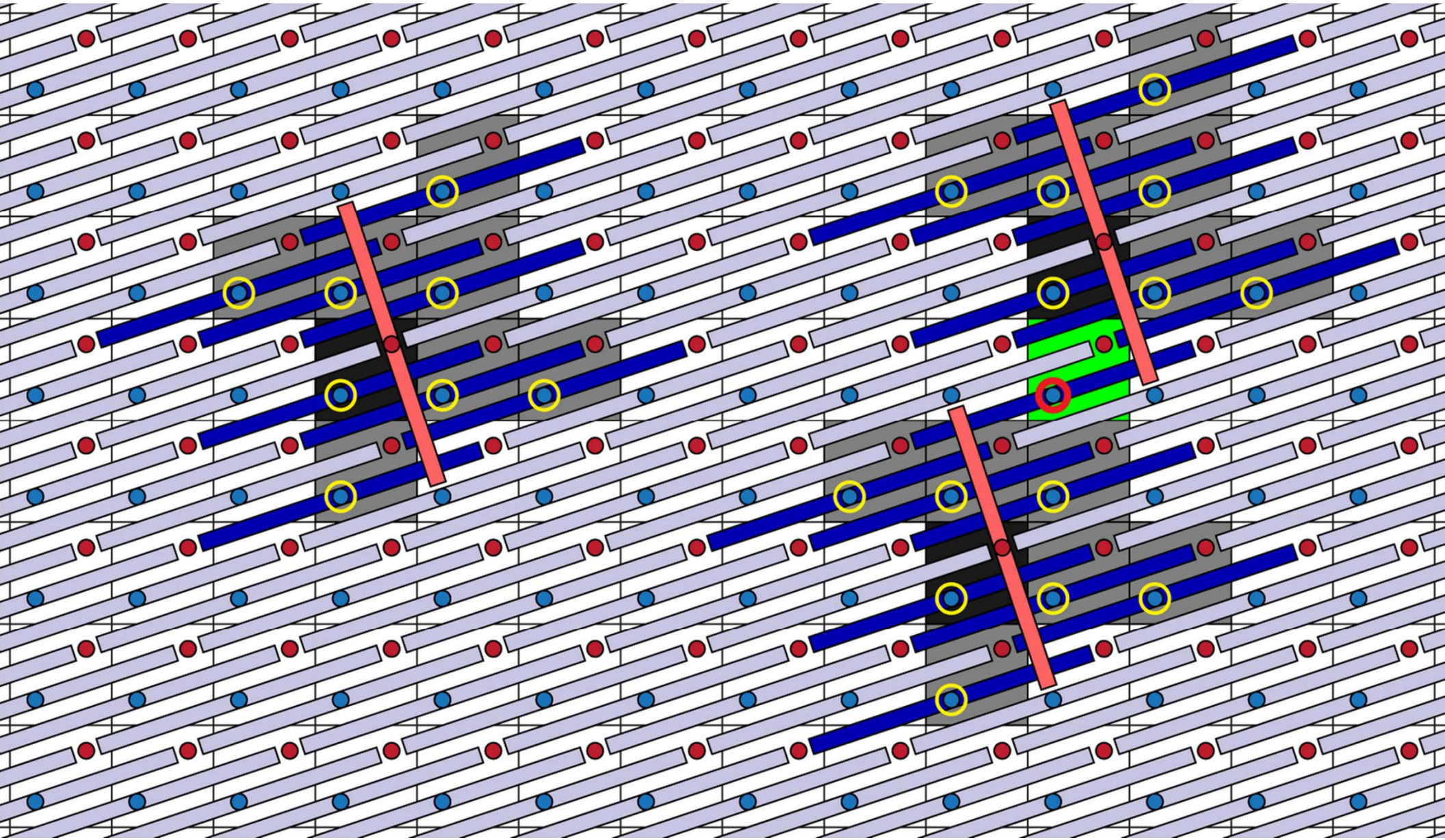
# Crossbar Construction



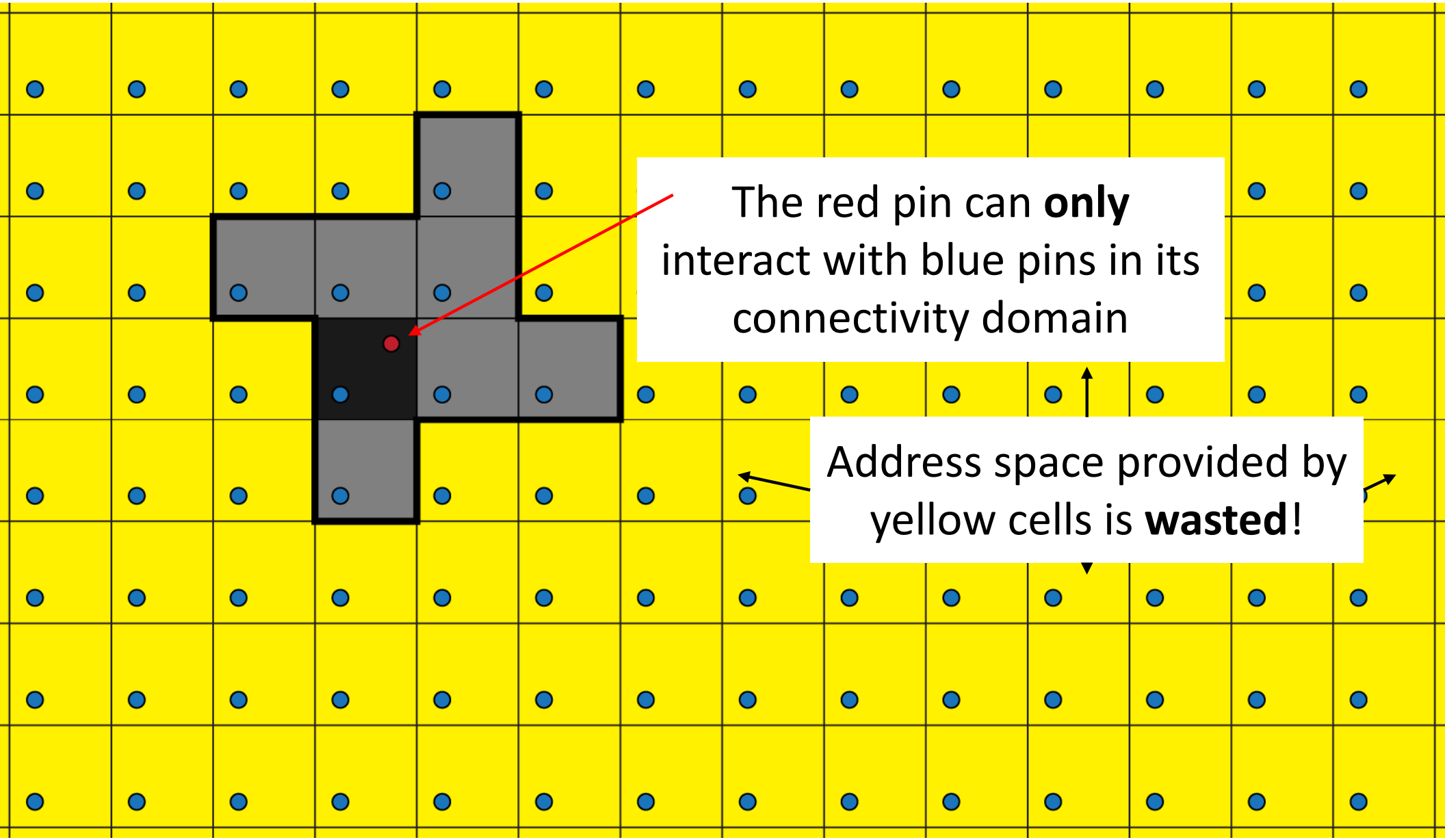
# Crossbar Construction



# Crossbar Construction



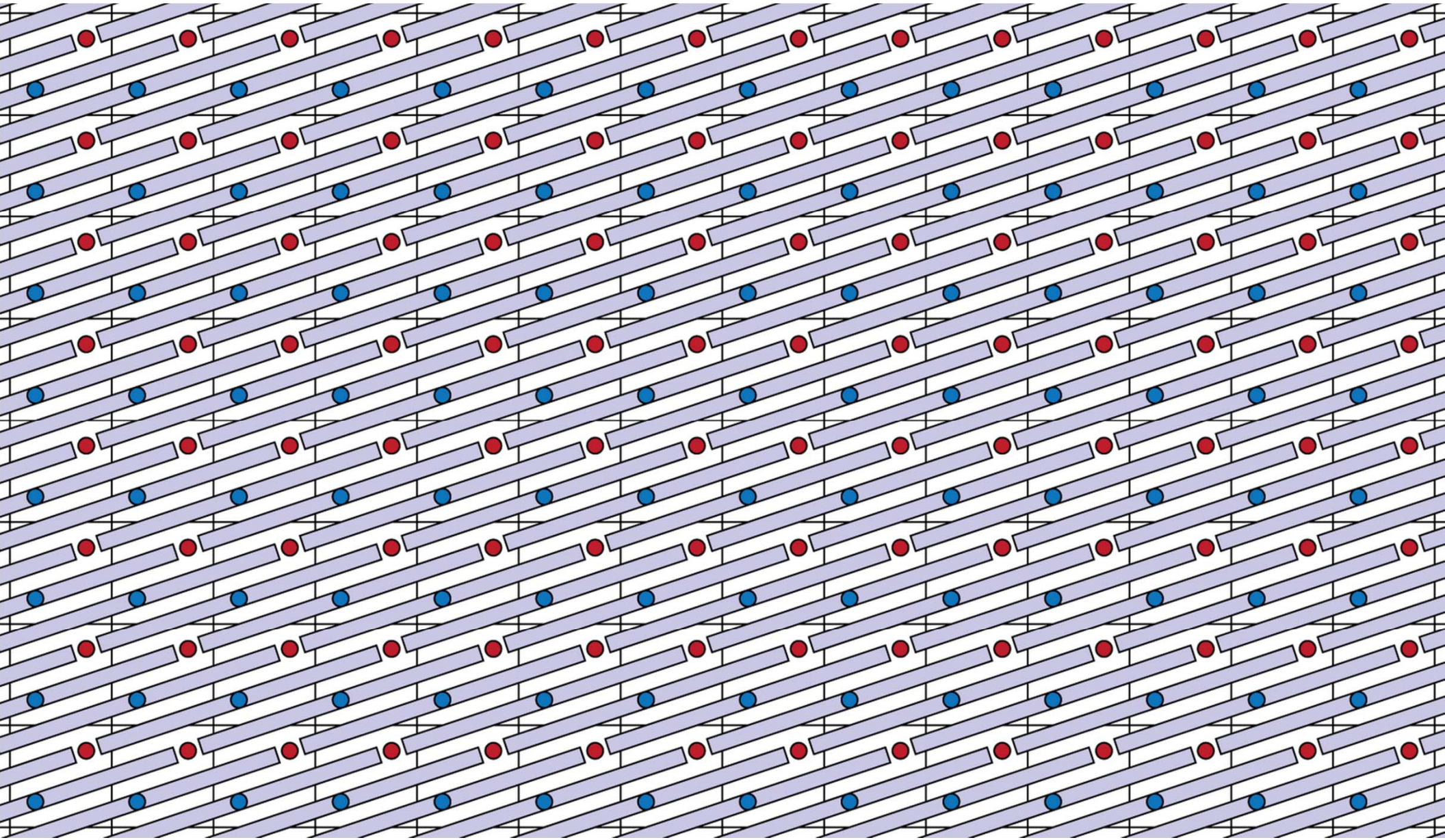
# Unused Address Space



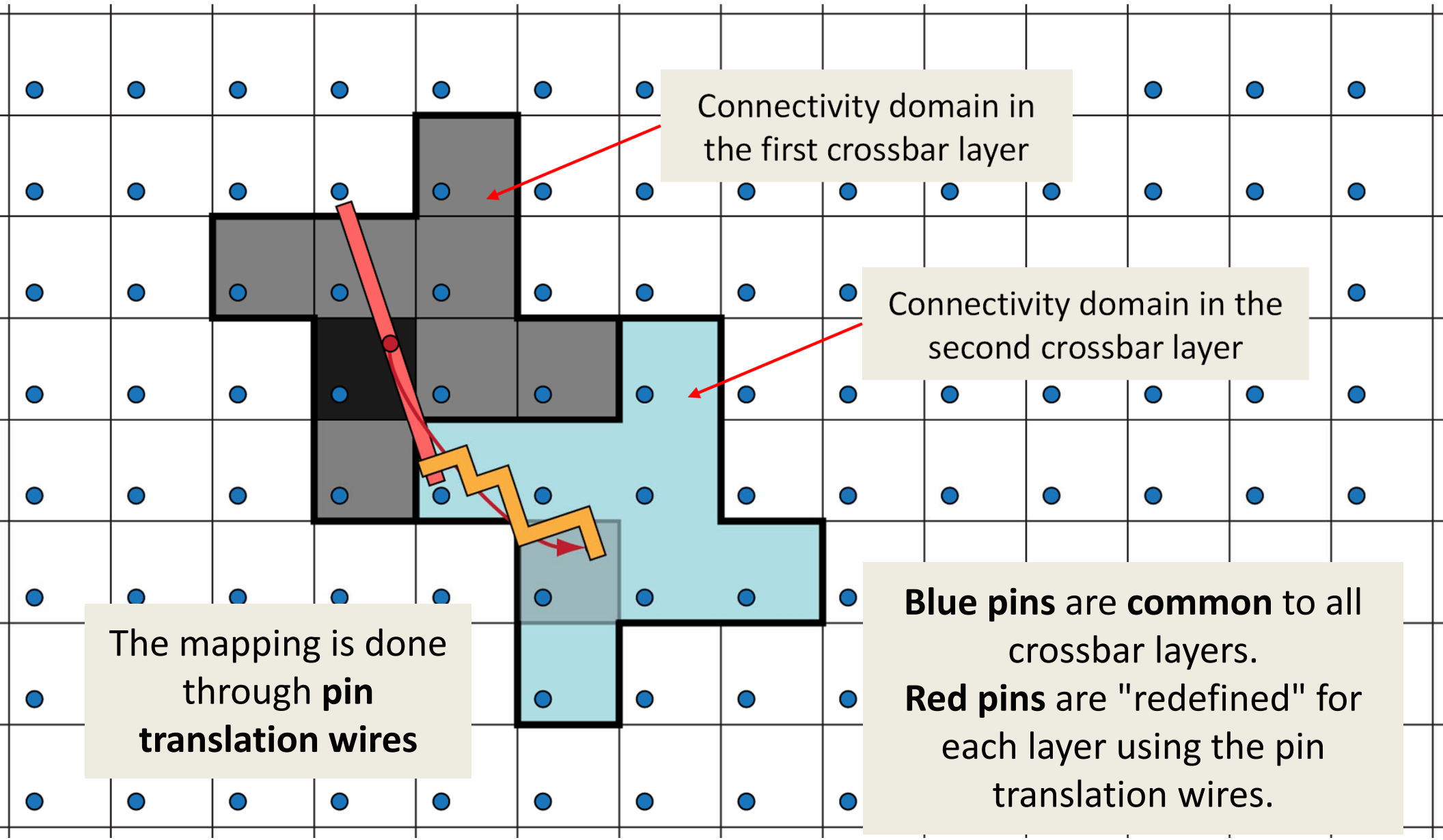
# Key Geometric Parameters

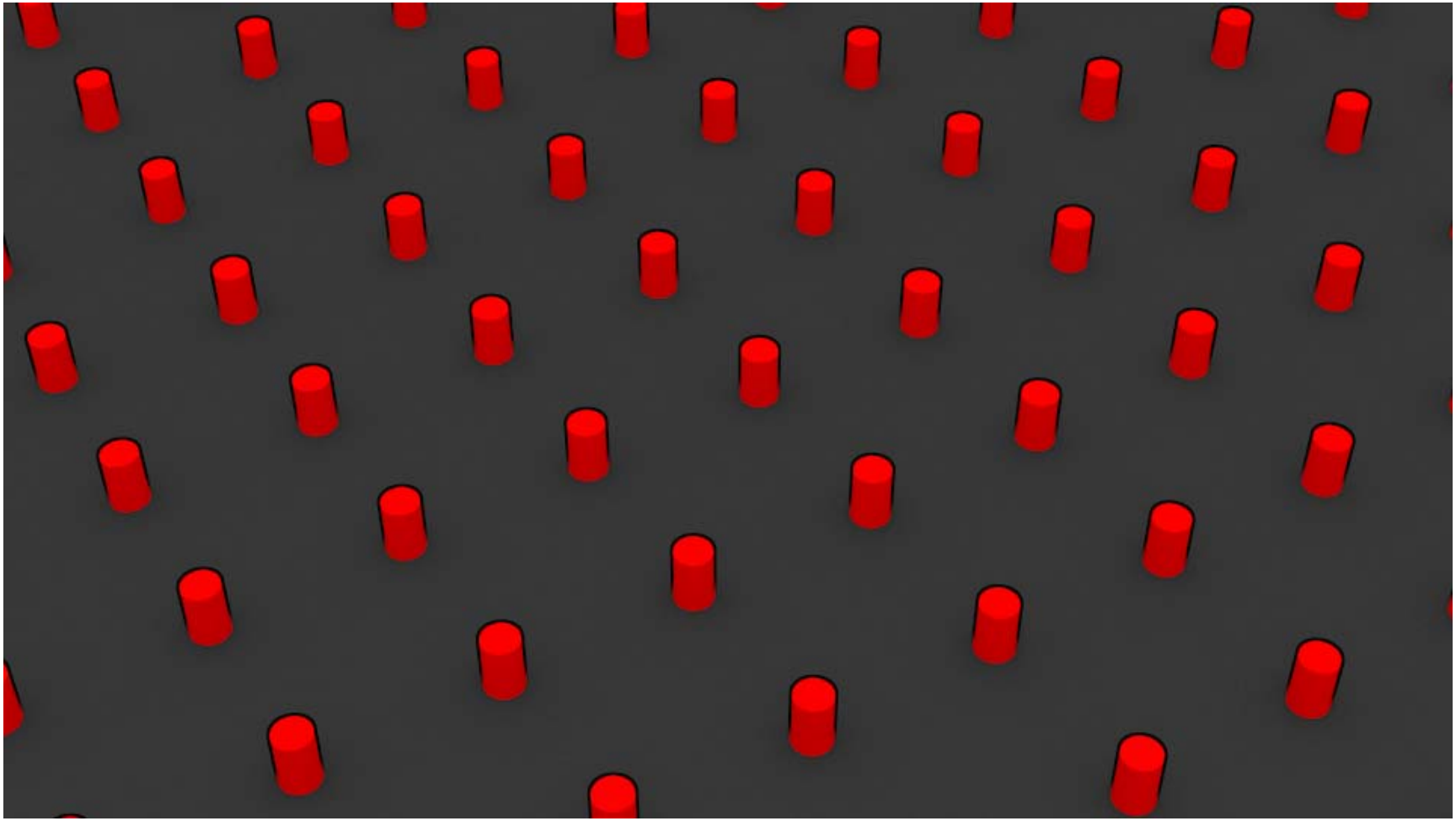
- Distance between nanowires is $2F_{NANO}$
- Size of cell is $2\beta F_{CMOS}$
- $\beta^2 = r^2 + 1$ where r is an odd integer > 1.
- Crossbar is tilted by an angle $\alpha$ equal to ArcTan(1/r) with respect to the pins.
- # of reachable crosspoints per wire segment is $r^2 - 1$

# Crossbar Construction – Bottom Level

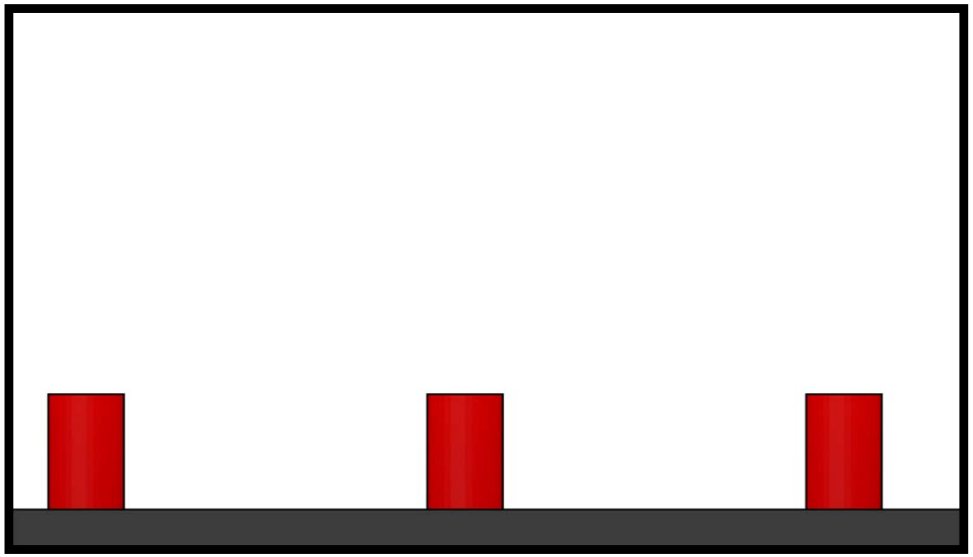


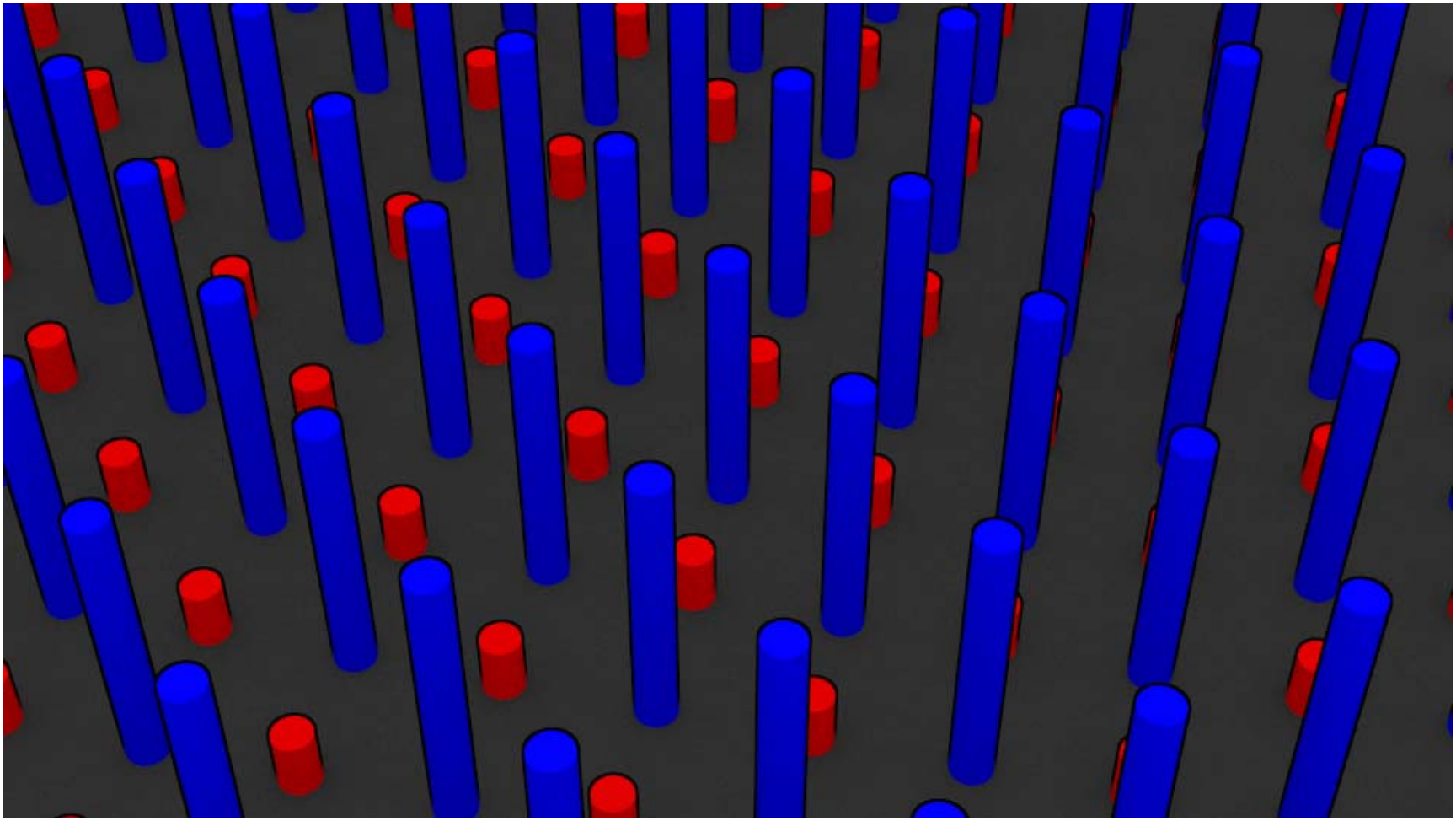
# Adding a Second Crossbar Layer



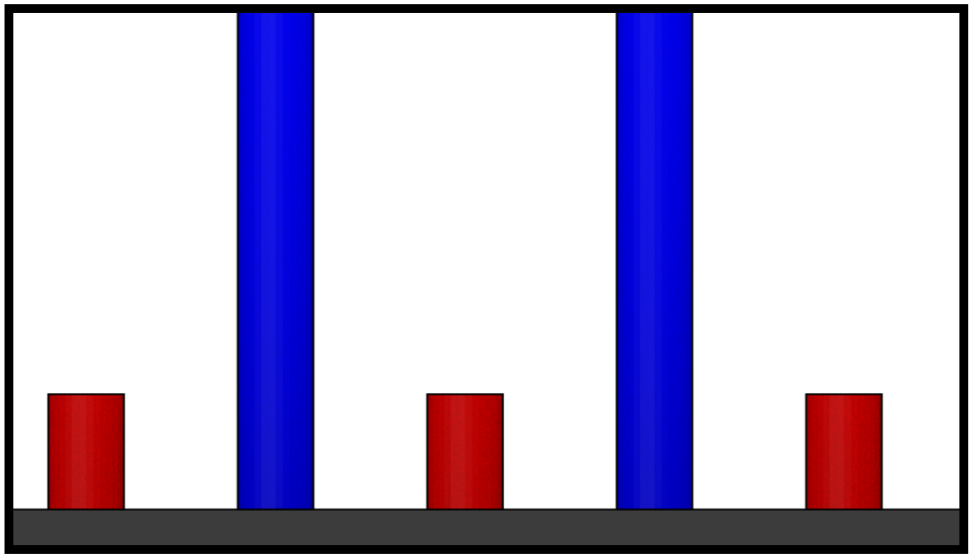


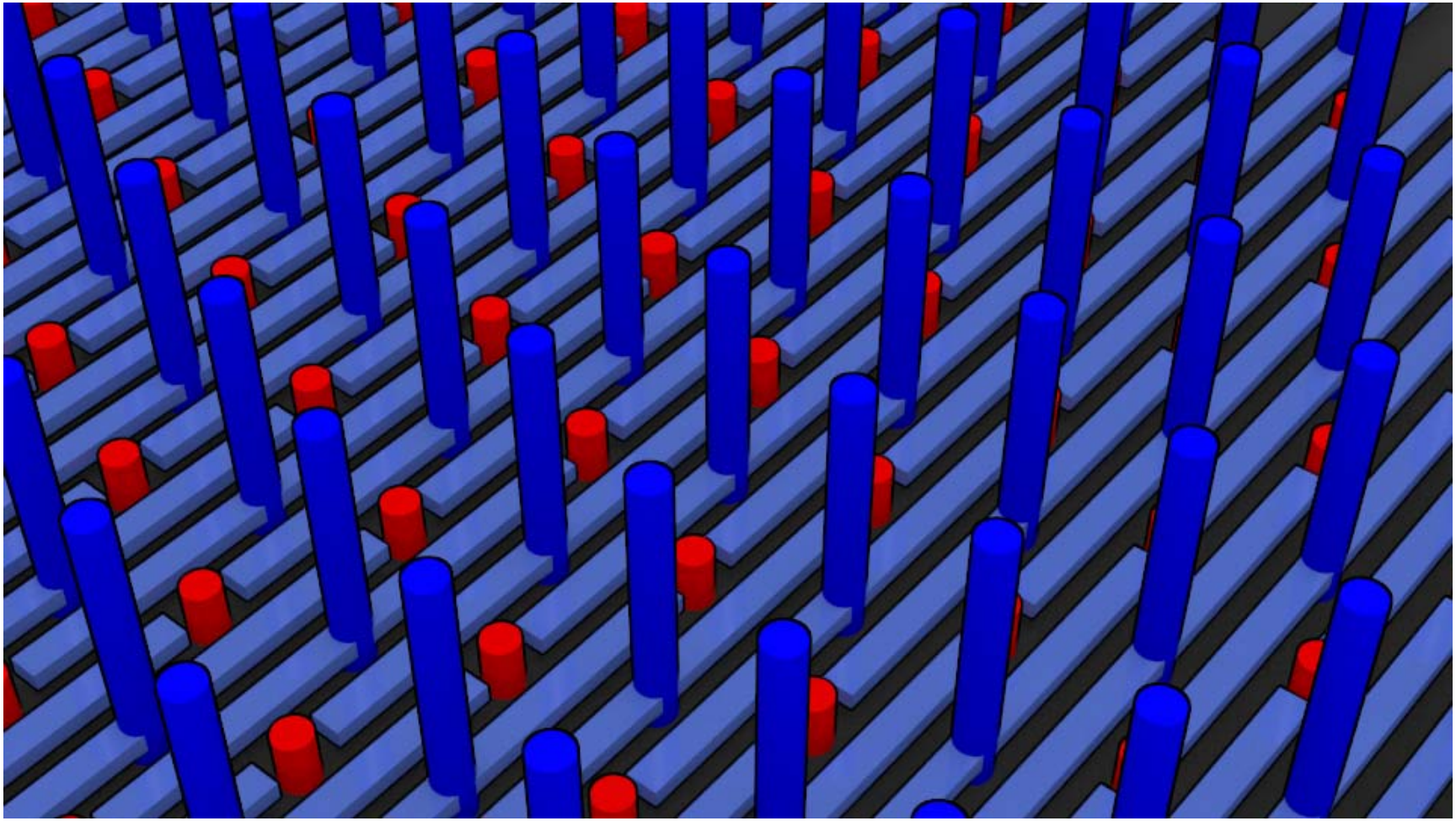
**First layer of red pins.**



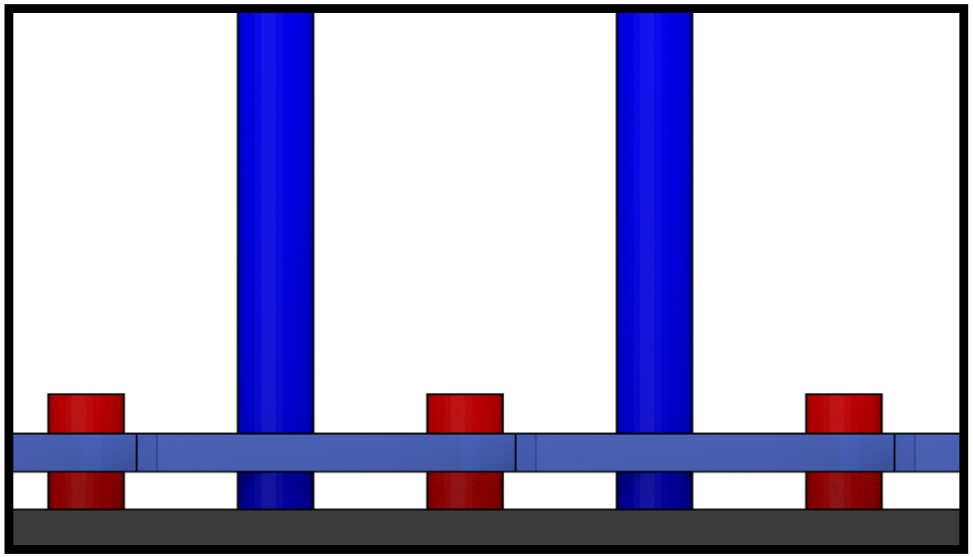


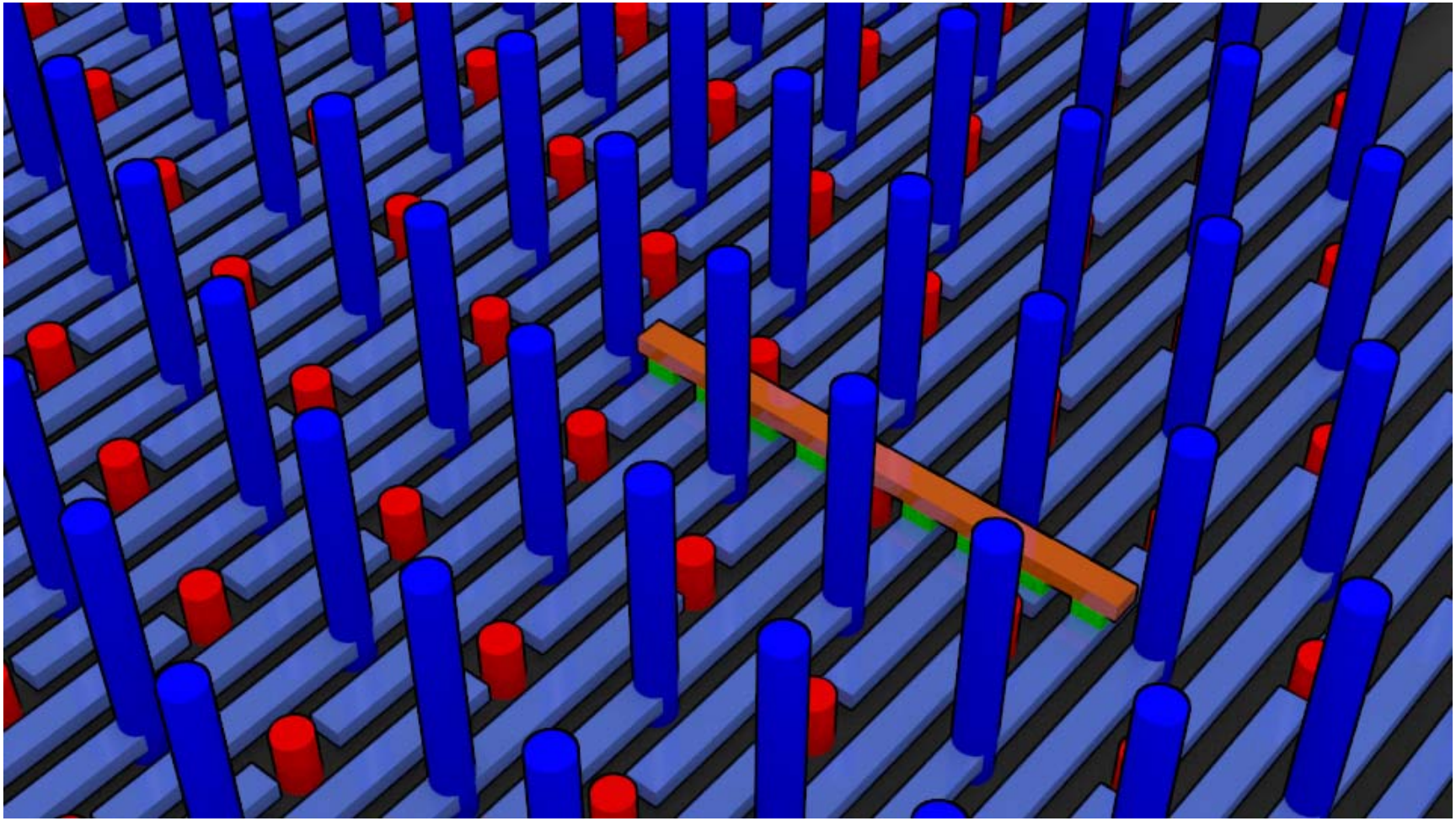
**First layer of red and blue pins.**



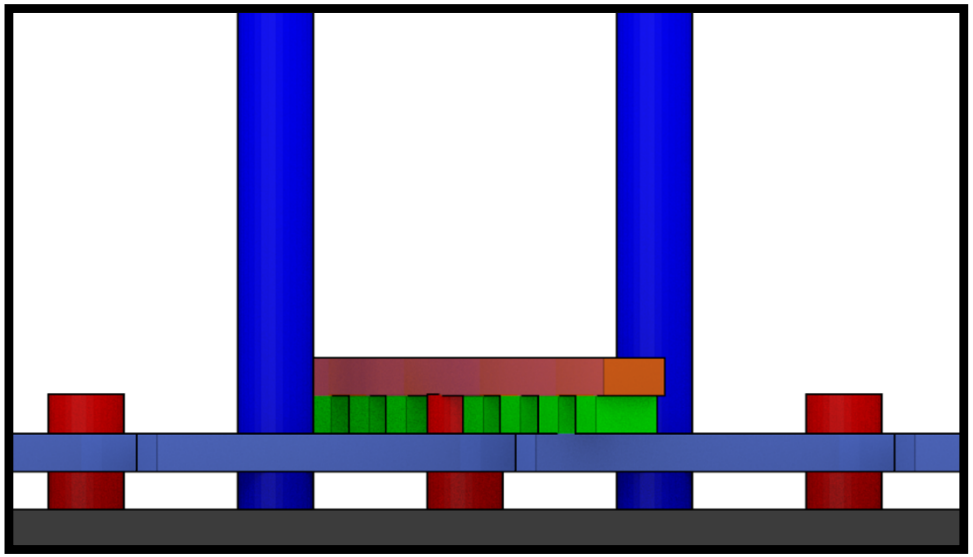


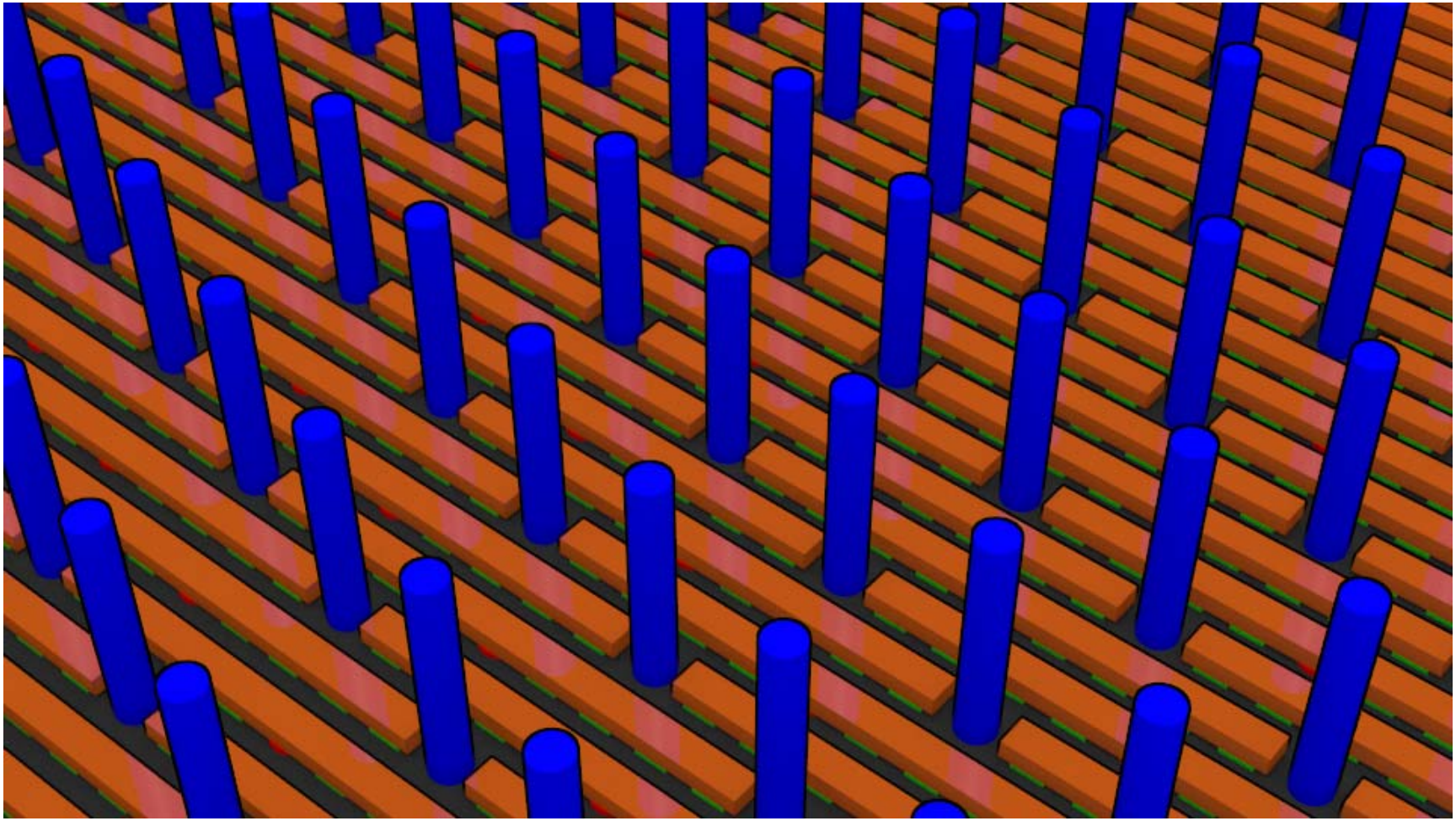
**Layer of (bluish) wires connected to the blue pins.**



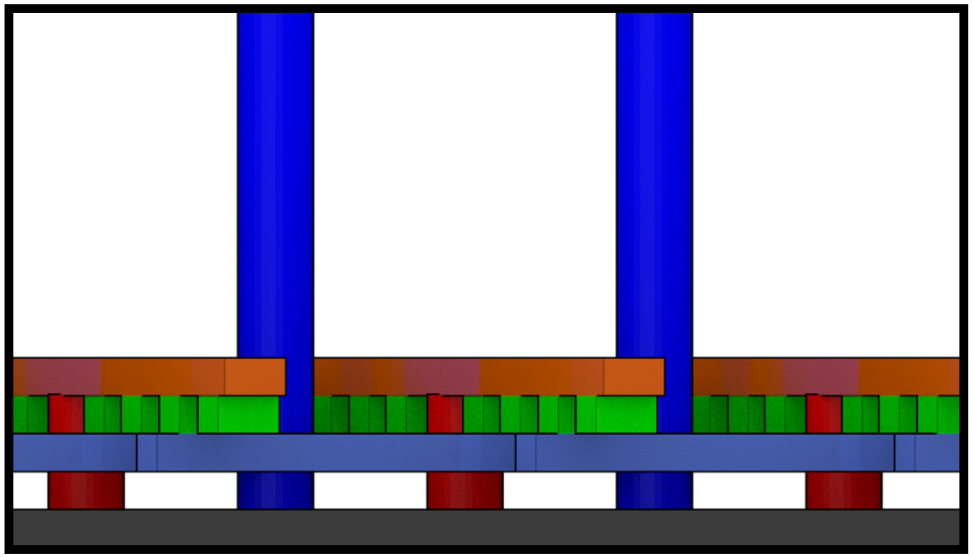


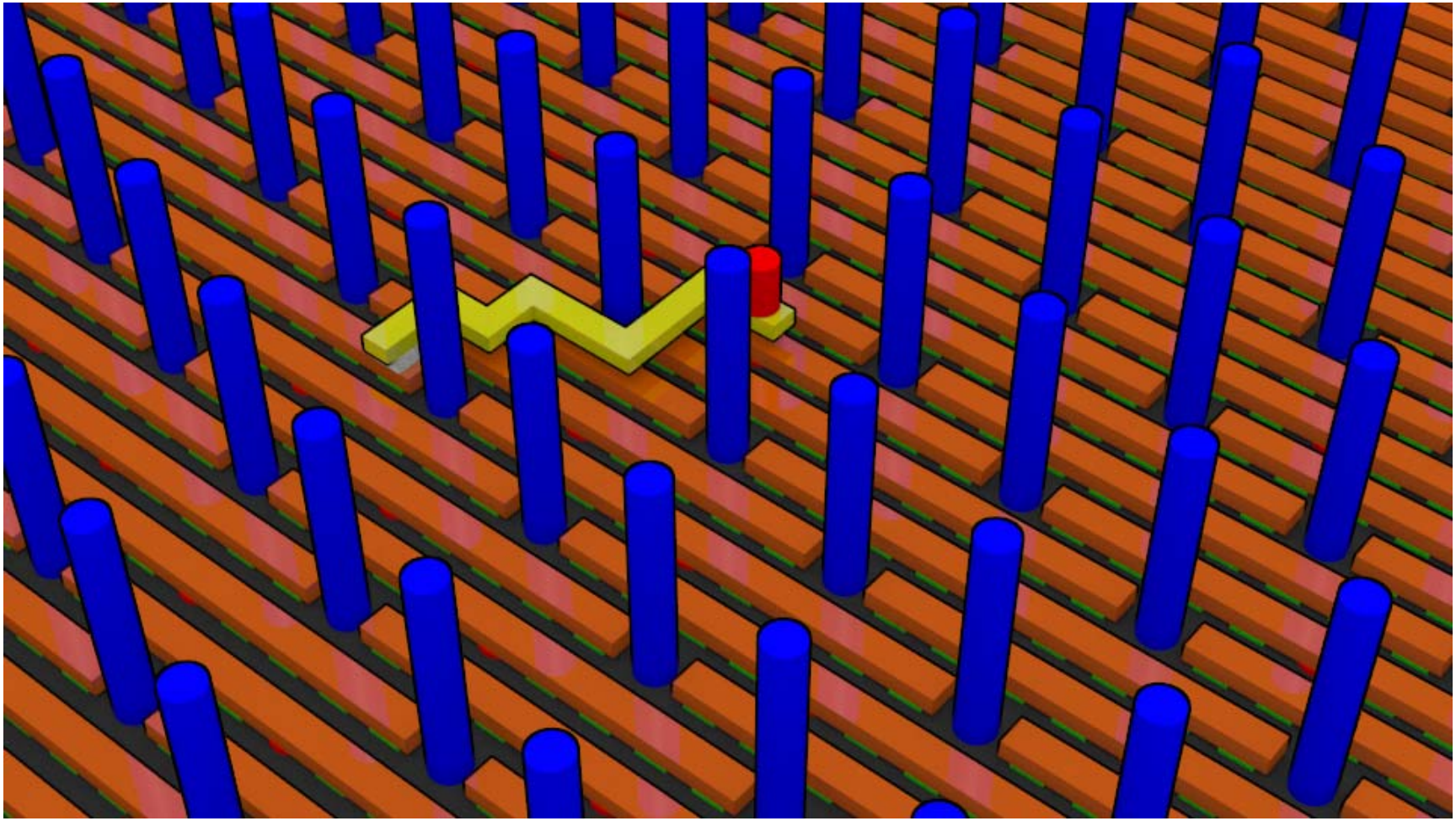
**Single (orange) wire connected to a red pin. The cross-points with the bottom wires are shown in green.**



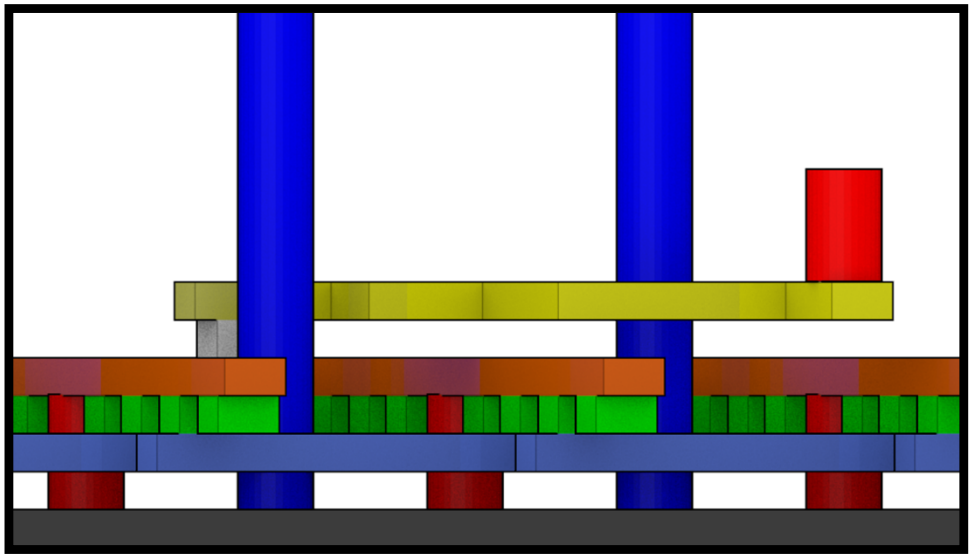


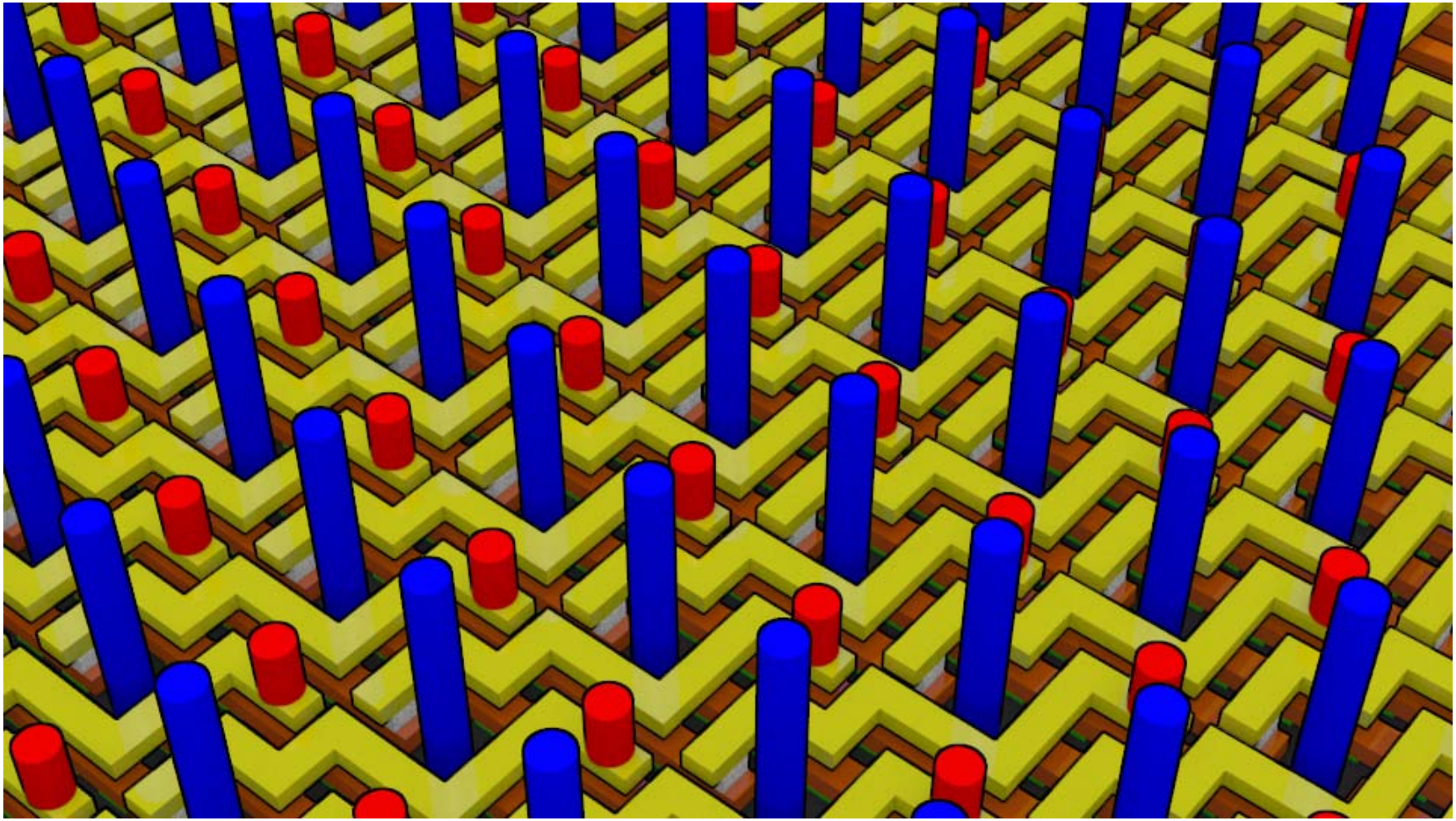
**First complete crossbar layer.**



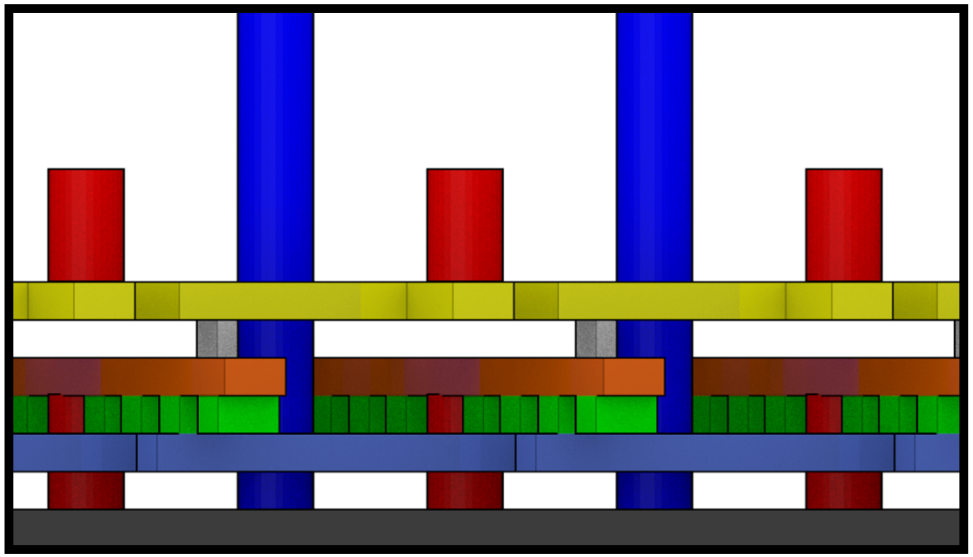


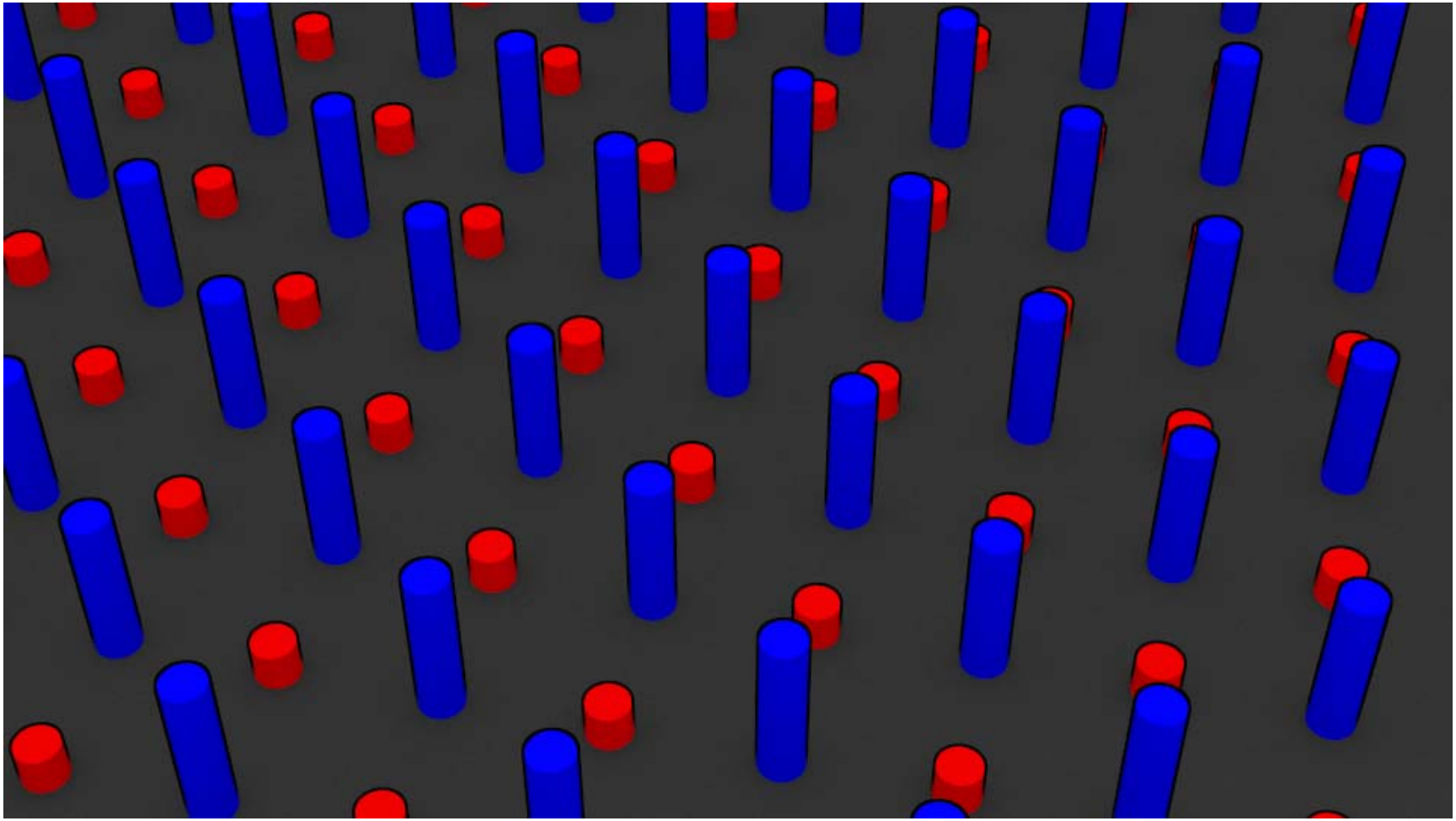
**A single pin translation wire (in yellow).**



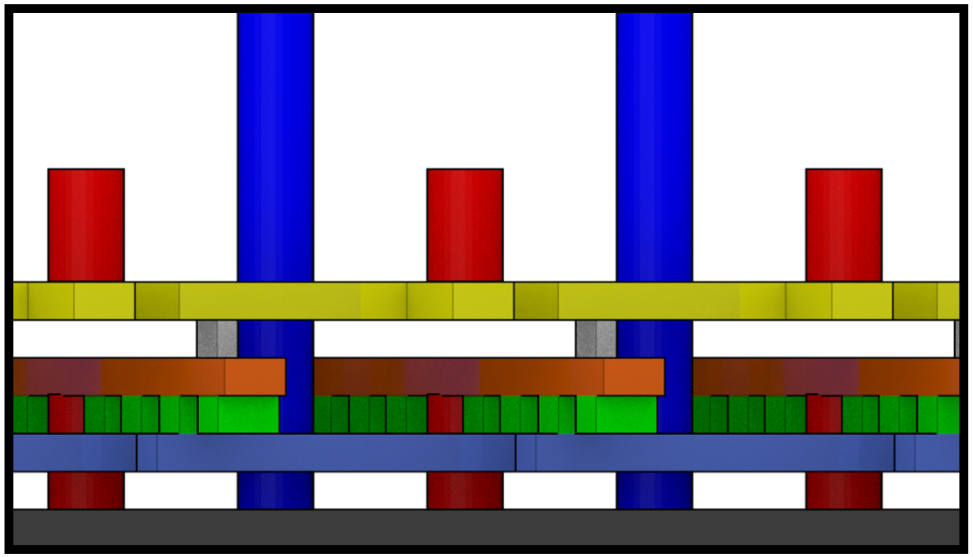


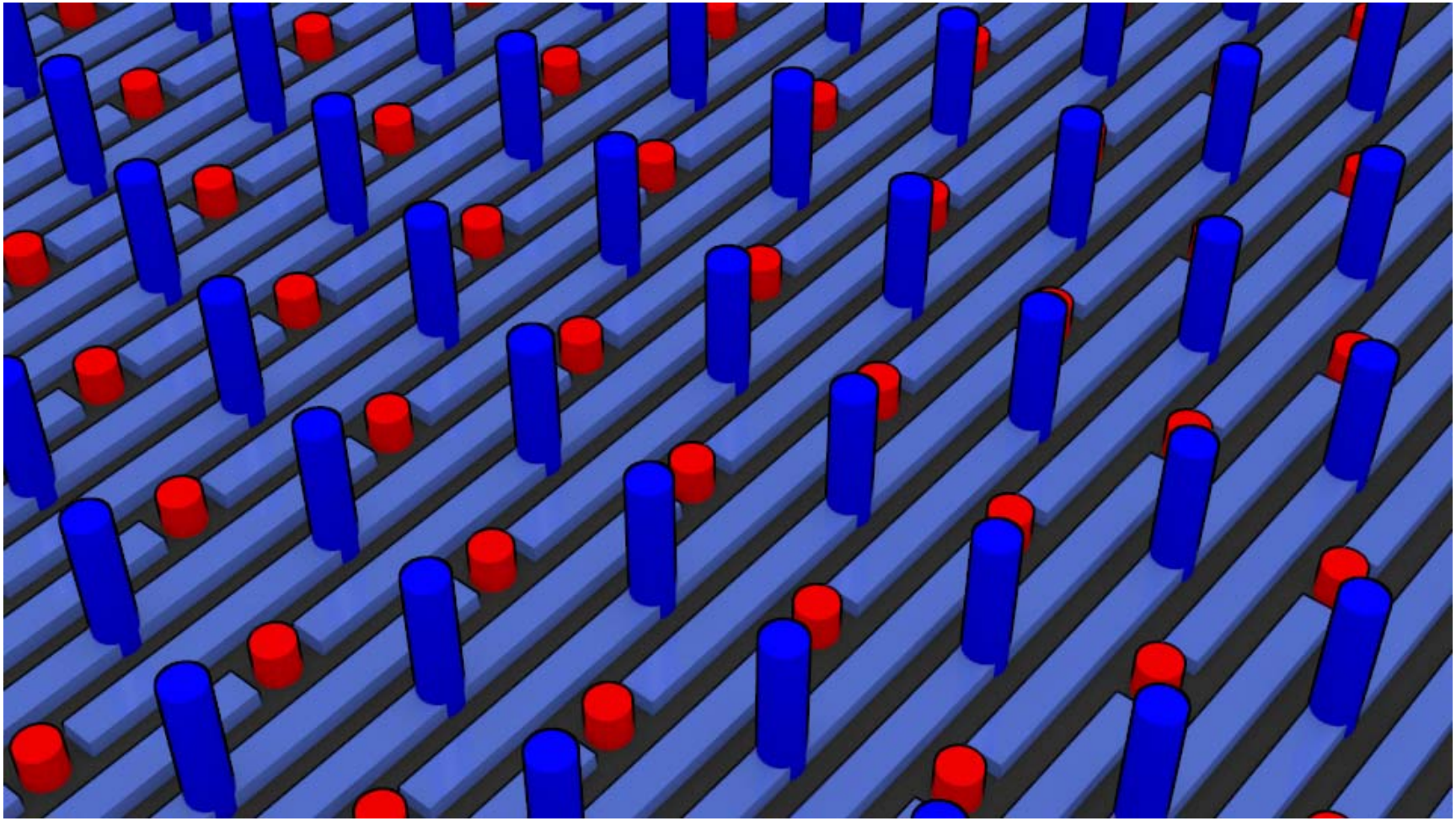
**Every orange wire is “translated” into another point using the same type of pin translation wire.**



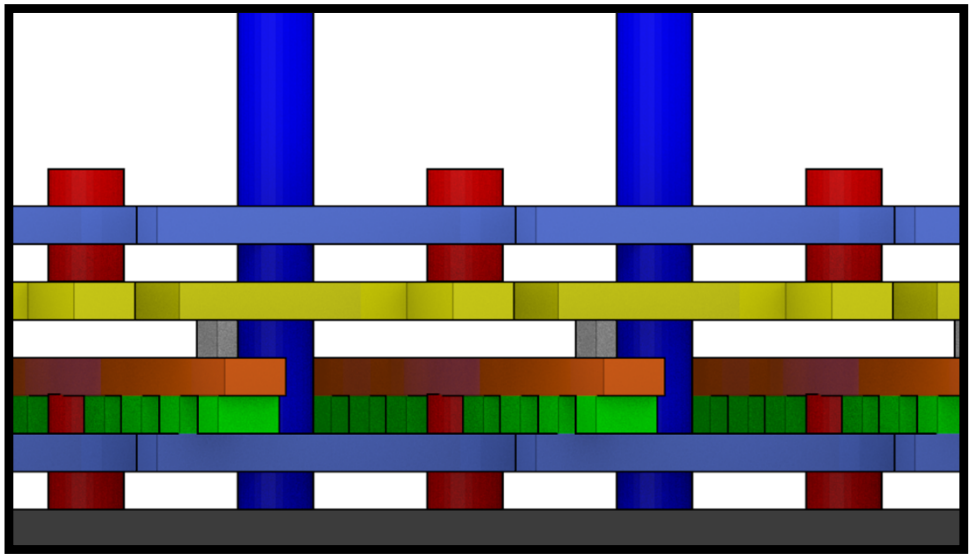


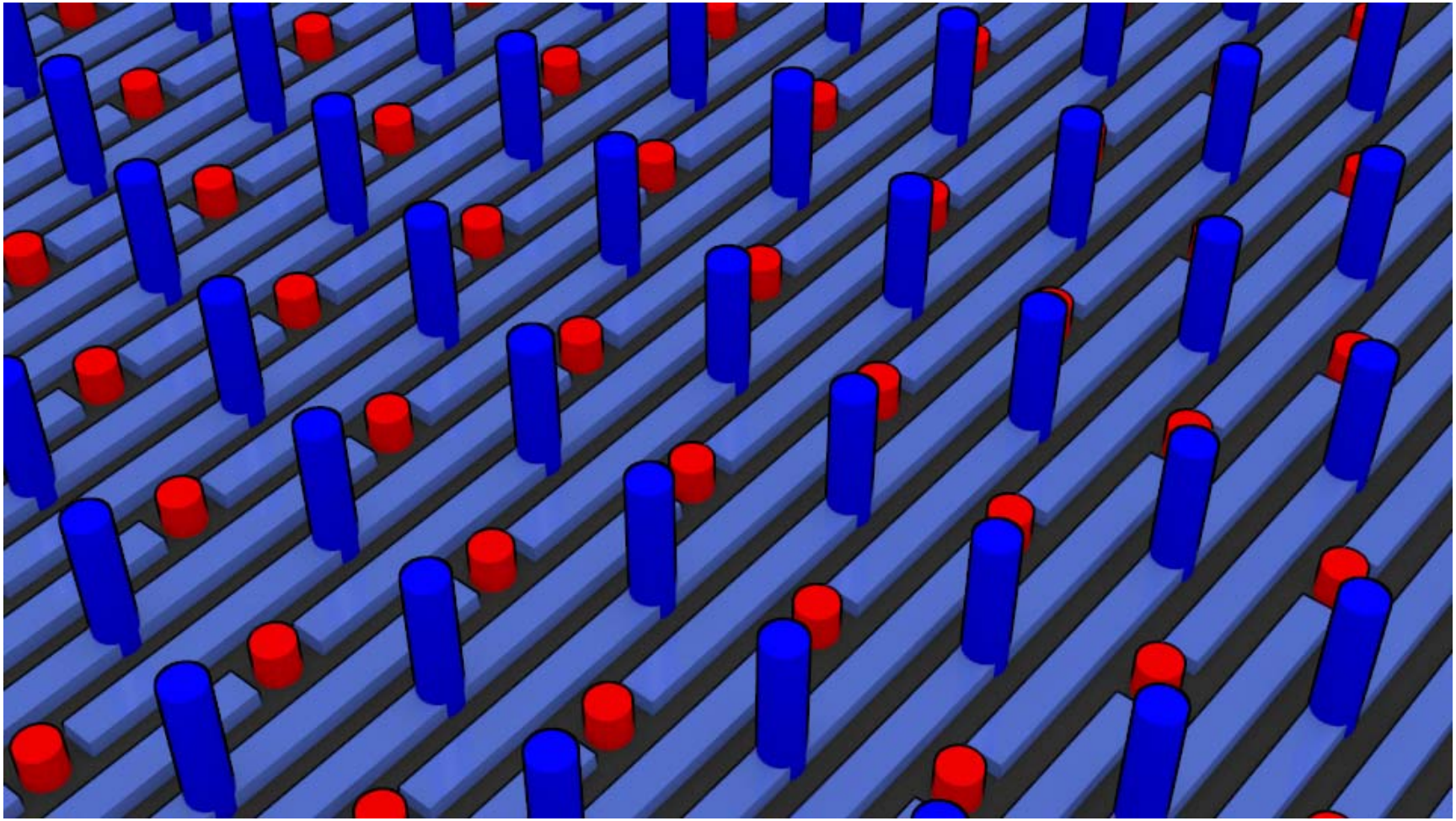
**The first crossbar layer with its pins translation wires are then “buried” in $SiO_2$**



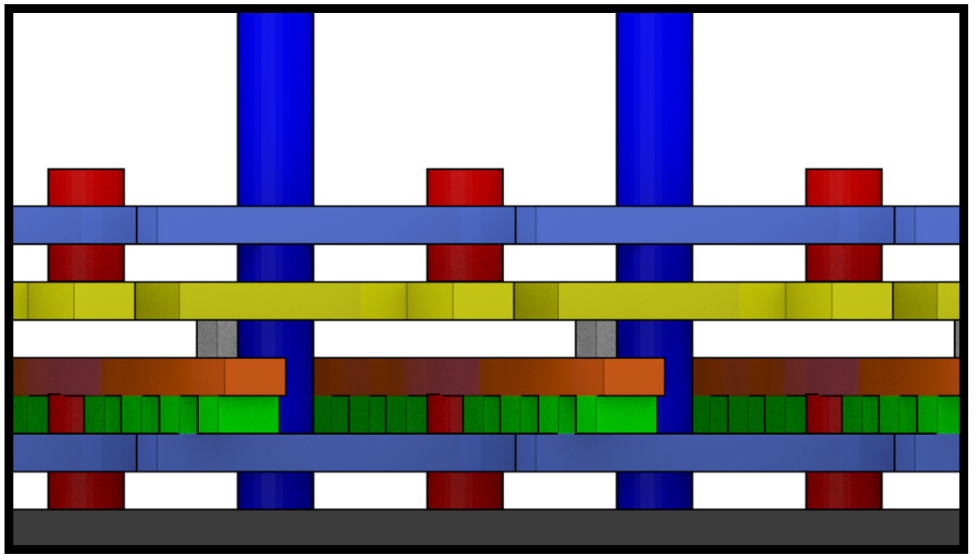


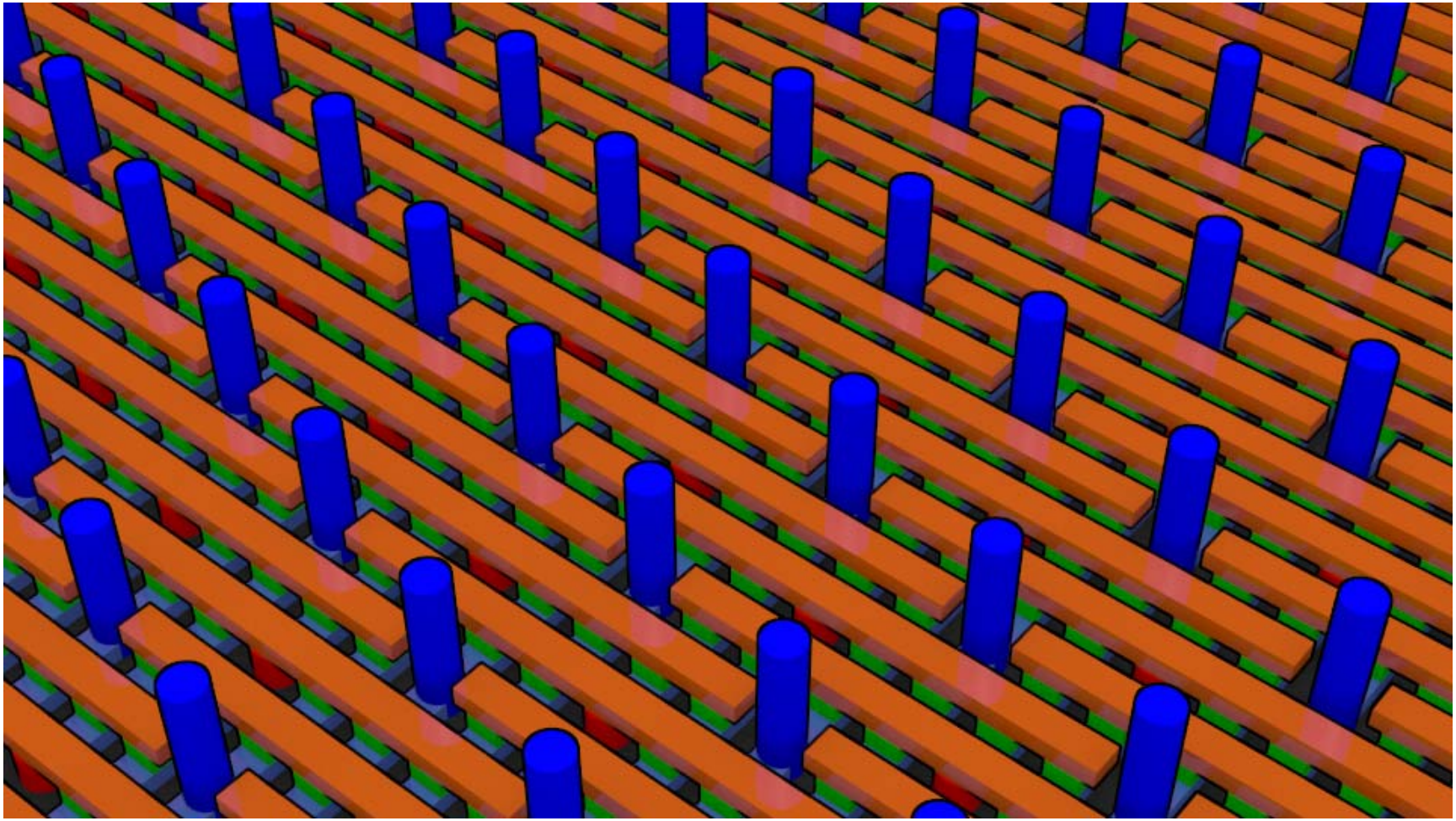
**We start to build the next crossbar layer (bluish wires)**



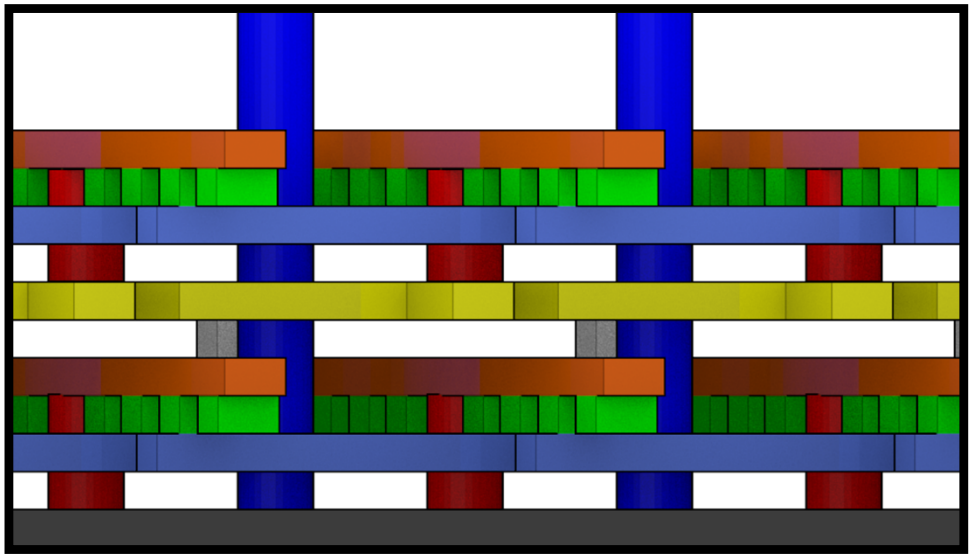


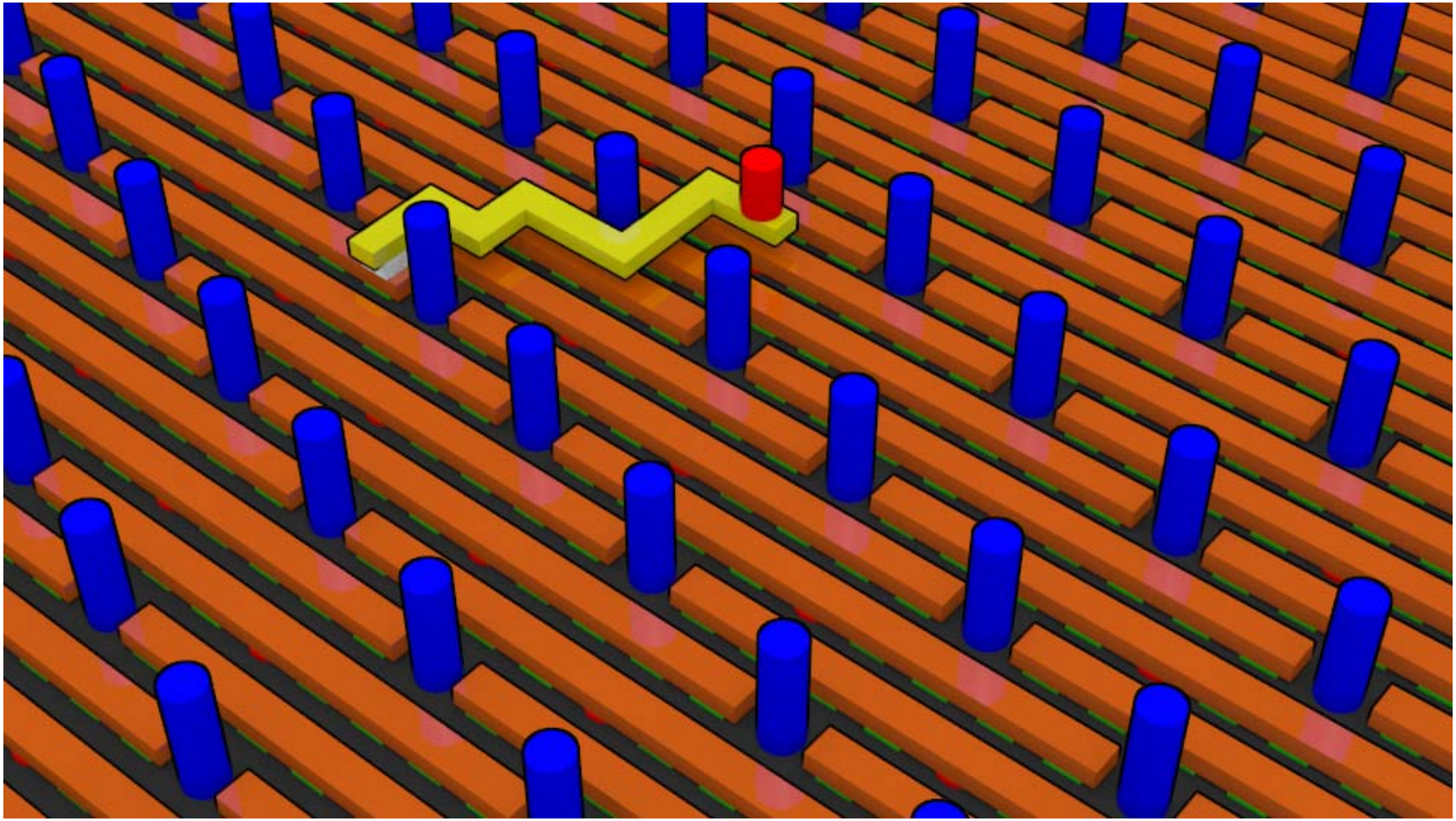
**We start to build the next crossbar layer (bluish wires)**



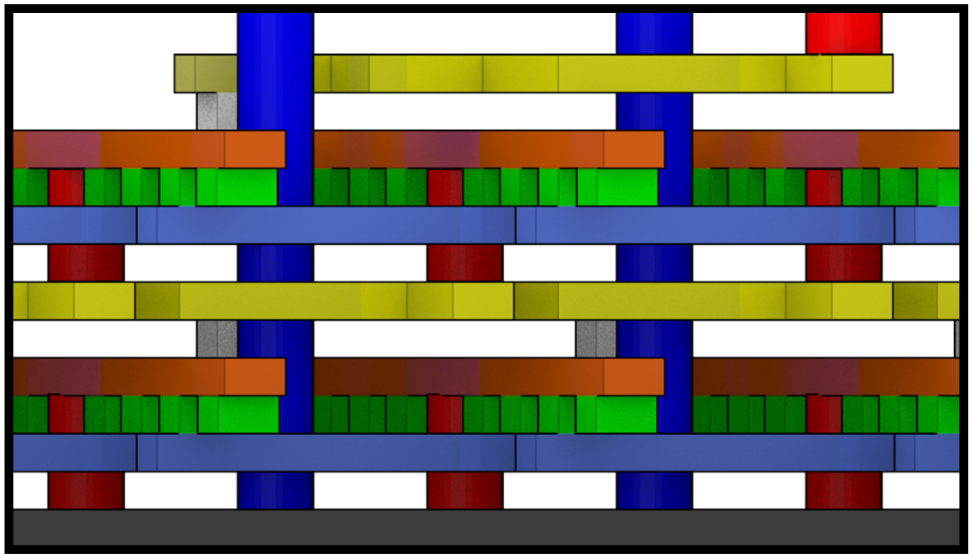


We add the orange wires (the cross-points are formed)





**And we add the pins translation wires and repeat the process…**



# Maximum Number of Layers

- Each layer has $N^2$ cells.
- There are $r^2 - 1$ cross points per cell.
- That gives us a total of $N^2(r^2 - 1)$ cross points per layer.
- The double decoding scheme allows us to address up to $N^4$ locations
- Which means that we can (potentially) have up to $N^2/(r^2 - 1)$ crossbar layers.

# How Does it Stand Up as a Memory?

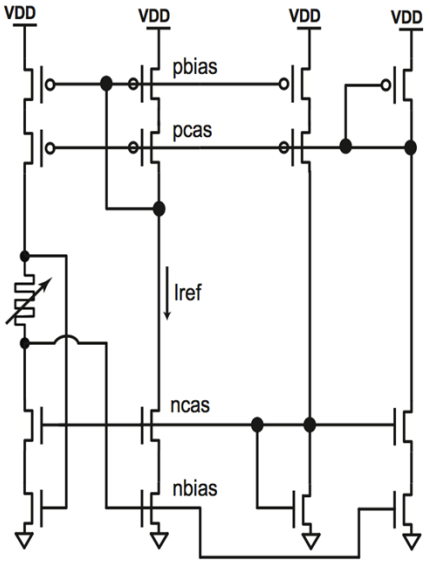
| | Memristor | PCM | STTRAM | DRAM | Flash | HDD |
|---|---|---|---|---|---|---|
| Density ($F^2$) | <4 | 8–16 | 37–64 | 6–8 | 4–6 | 2/3 |
| Energy per bit† (pJ) | 0.1–3 | 2–27 | 0.1 | 2 | 10000 | 1–10x$10^9$ |
| Read time (ns) | 10-100(?) | 20–70 | 10–30 | 10–50 | 25000 | 5–8x$10^6$ |
| Write time (ns) | ~10 | 50–500 | 13–95 | 10–50 | 200000 | 5–8x$10^6$ |
| Retention | years | years | weeks? | <<second | years | years |
| Endurance (cycles) | >$10^{12}$ | $10^7$ | $10^{15}$ | $10^{15}$ | $10^6$ | $10^4$ |

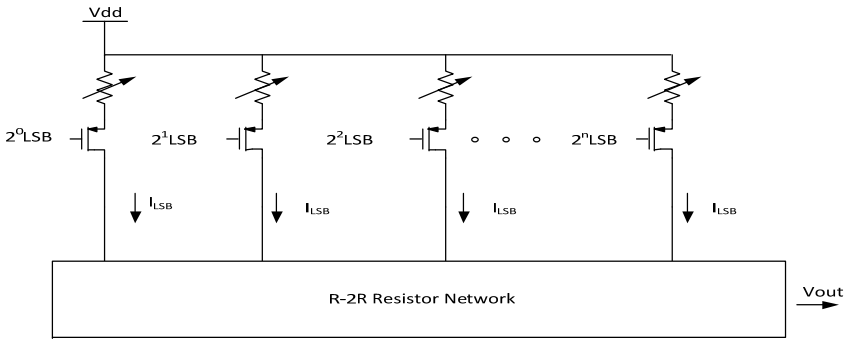
# If Successful, 3D Hybrids Can Achieve…..

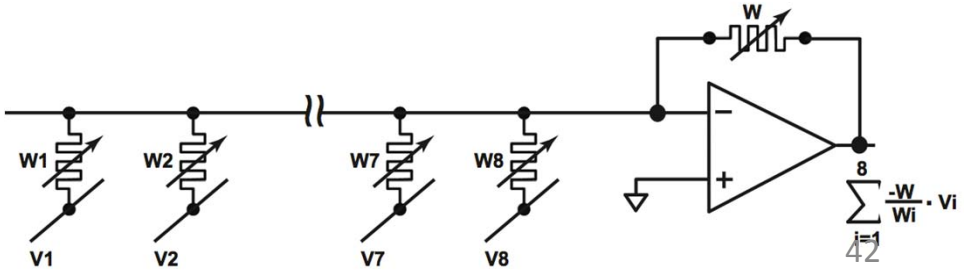
- Unprecedented memory density
  - Footprint of a nano-device is $4F_{nano}^2/K$, for K vertically integrated crossbar layers
  - Potentially up to $10^{14}$ bits on a single 1-$cm^2$ chip
- Enormous memory bandwidth
  - Potentially up to $10^{18}$ bits/second/$cm^2$
- At manageable power dissipation
- With abundant redundancy for yield/reliability

# Thrust Area #1: Application/Architecture/Ckt Exploration

- Memory arrays for high-performance computing
- **CMOL-based FPGA**
- Neuromorphic networks for bio-inspired information processing
- Evolvable analog circuits
- Tunable bias network for analog design
- Weighted multiply and add circuits
- High precision Digital-to-Analog converter

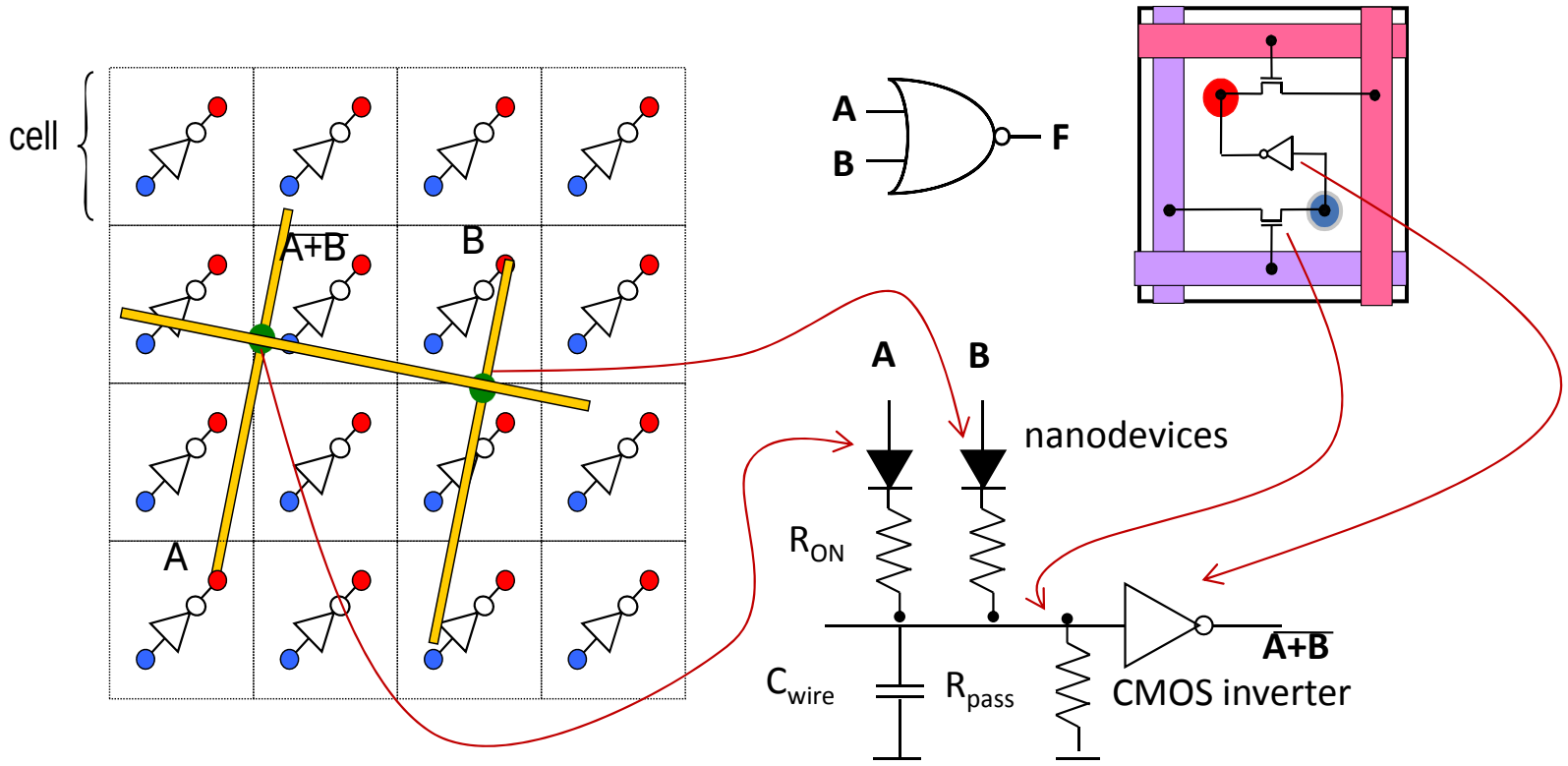


# CMOL-Based FPGA

- Programming for xpoint memristors similar to CMOL digital memories
- Uniform fabric with CMOS inverter cells
- Crossbar wires for routings



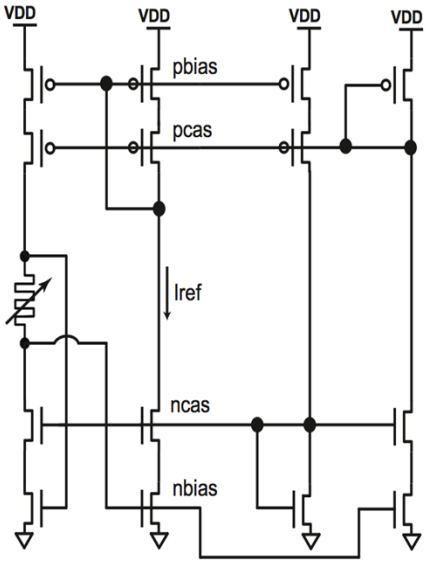
# Density: CMOS vs. CMOL

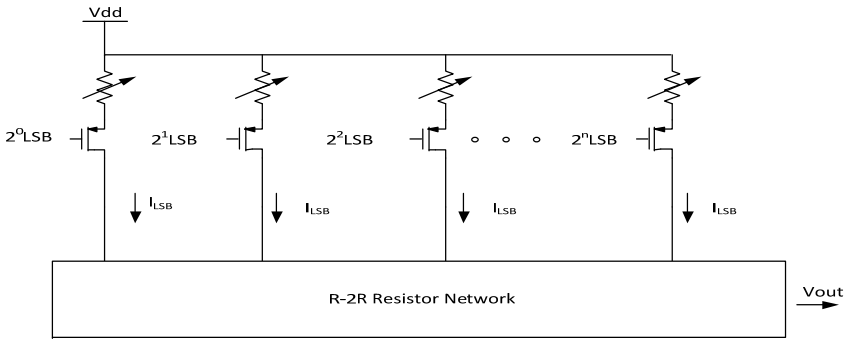
| Metrics (units) | 2009 | 2010 | 2011 | 2012 | 2013 | Comments |
|---|---|---|---|---|---|---|
| Half-pitch $F_{CMOS}$ (nm) | 50 | 45 | 40 | 36 | 32 | In accordance with ITRS |
| Half-pitch $F_{nano}$ (nm) | 20 | 18 | 16 | 14 | 12 | - |
| CMOS memories (Gbits/cm$^2$) | 6.7 | 8.2 | 10.5 | 13 | 16 | Follows ITRS (with $A = 6F^2_{CMOS}$) |
| CMOL memories (Gbits/cm$^2$) | 4 | 10 | 23 | 36 | 67 | Initial progress impacted by $q$ |
| CMOS FPGA (Mgates/cm$^2$) | 0.4 | 0.5 | 0.6 | 0.8 | 1.0 | Rescaled from 0.18 μm rules |
| CMOL FPGA (Mgates/cm$^2$) | 625 | 775 | 1,000 | 1,200 | 1,500 | - |

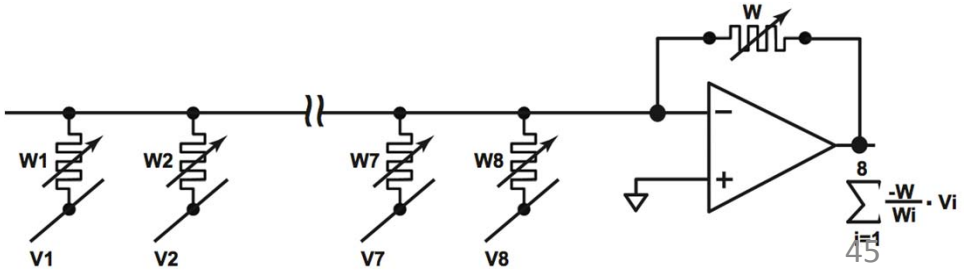
| Metrics (units) | 2016 | 2019 | 2022 | 2025 | 2028 | Comments |
|---|---|---|---|---|---|---|
| Half-pitch $F_{CMOS}$ (nm) | 30 | 28 | 26 | 24 | 22 | Grows slower than in ITRS |
| Half-pitch $F_{nano}$ (nm) | 10 | 6 | 4 | 3.5 | 3 | - |
| CMOS memories (Gbits/cm$^2$) | 18 | 21 | 25 | 29 | 35 | Follows $A = 6F^2_{CMOS}$ |
| CMOL memories (Gbits/cm$^2$) | 100 | 350 | 900 | 1,200 | 1,700 | Spectacular progress at lower $q$ |
| CMOS FPGA (Mgates/cm$^2$) | 1.1 | 1.3 | 1.5 | 1.7 | 2.1 | Rescaled from 0.18 μm rules |
| CMOL FPGA (Mgates/cm$^2$) | 1,700 | 2,000 | 2,300 | 2,700 | 3,200 | - |

# Thrust Area #1: Application/Architecture/Ckt Exploration

- Memory arrays for high-performance computing
- CMOL-based FPGA
- Neuromorphic networks for bio-inspired information processing
- Evolvable analog circuits
- Tunable bias network for analog design
- Weighted multiply and add circuits
- High precision Digital-to-Analog converter



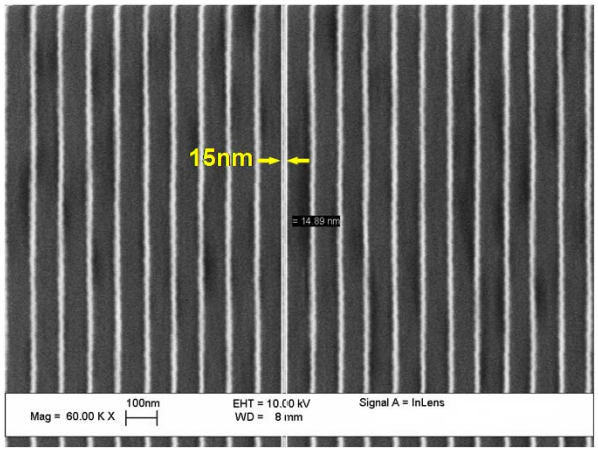
# Thrust Areas:
# # 2: High-Performance/-Yield Devices
# # 3: 3D Hybrids Integration

**Integrating CMOS with devices of different materials:**

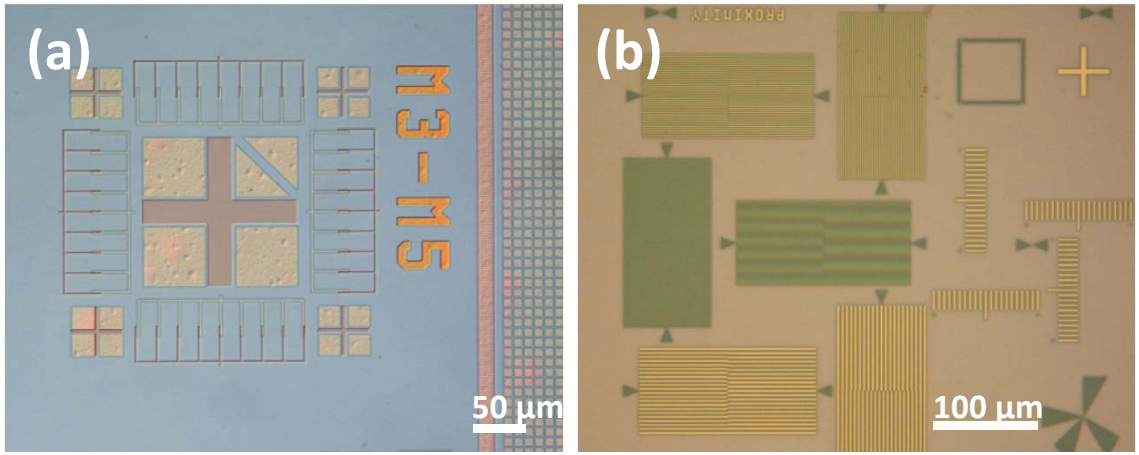
- a-Si
- Metal oxide
- Organic
- Solid-state electrolyte

**Using:**

- Nanoimprint
- E-beam lithography
- Optical lithography
- Heterogeneous wafer-level integration



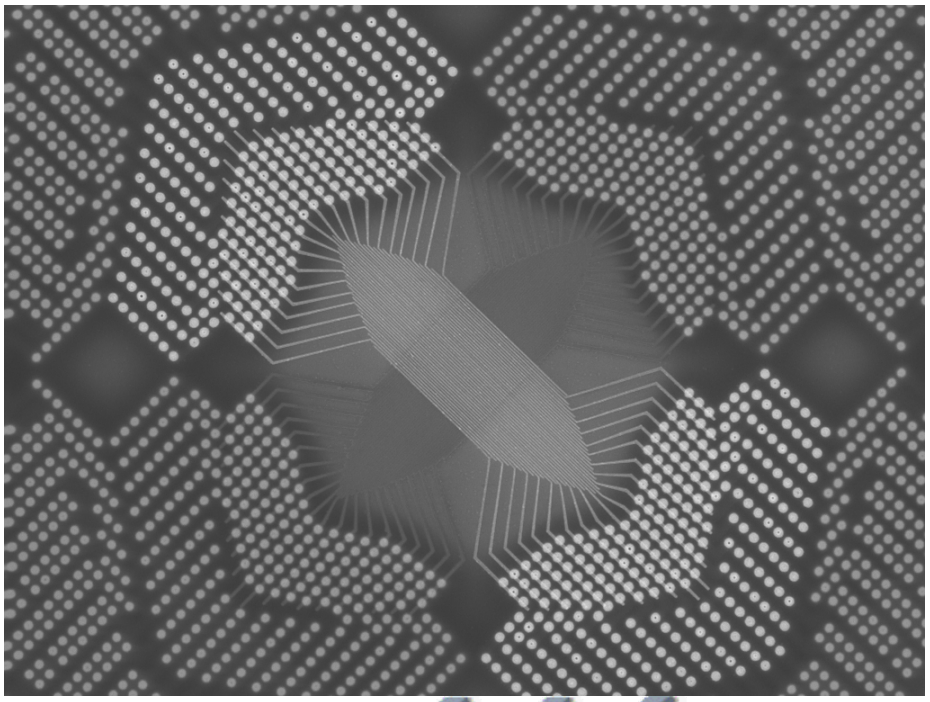

E-Beam Crossbar Arrays (Lu)

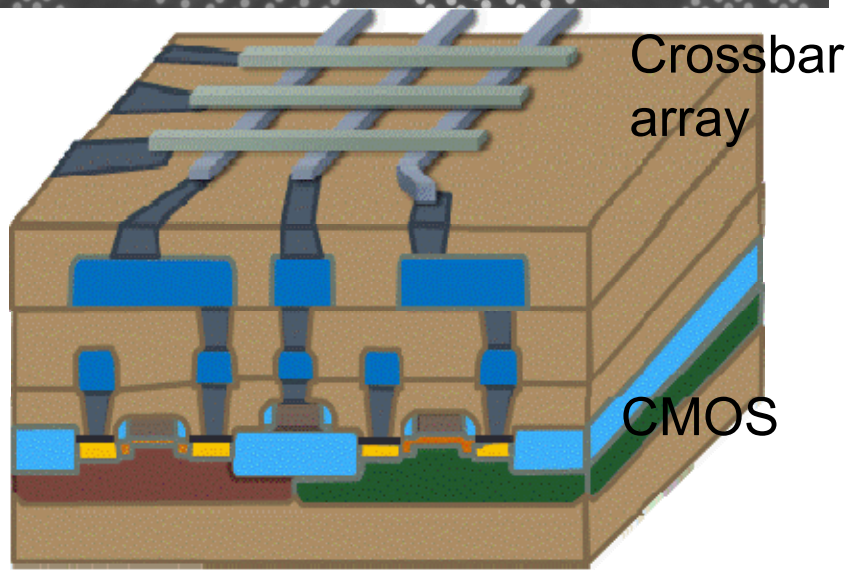


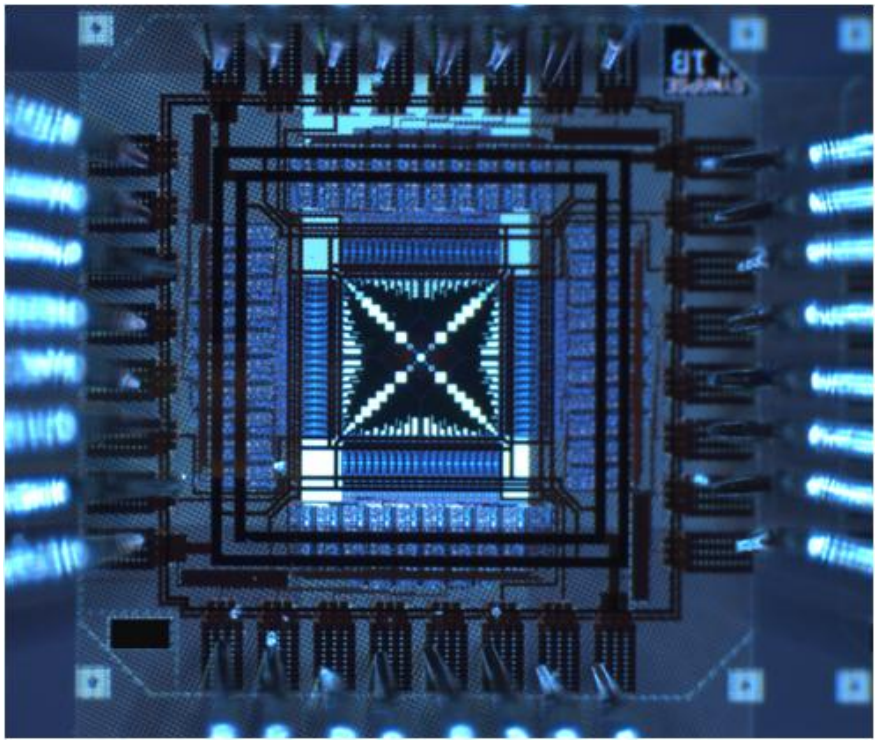

<20nm Overlay Alignment (Xia)

# Integrated Crossbar Array/CMOS System

PI: Lu



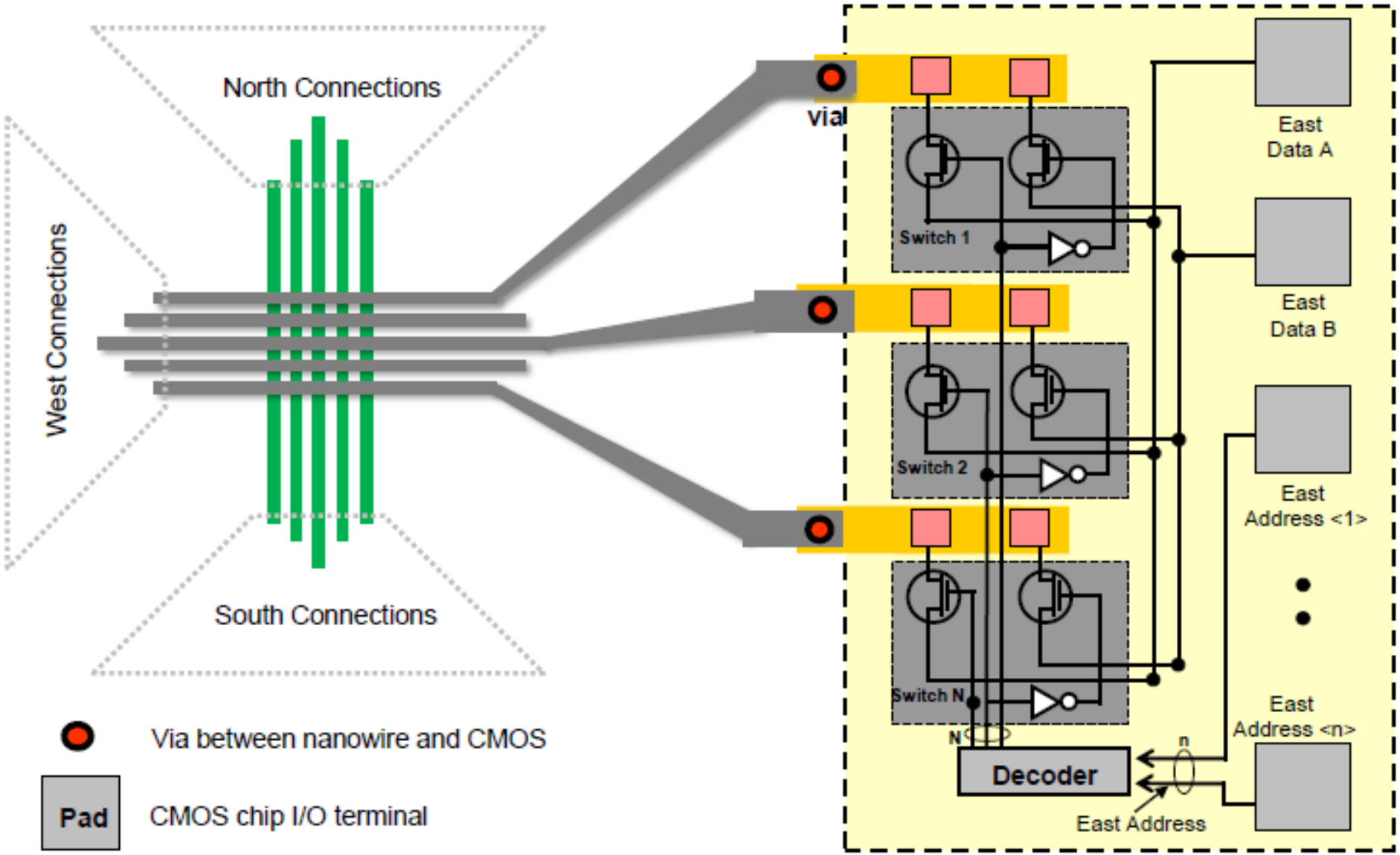


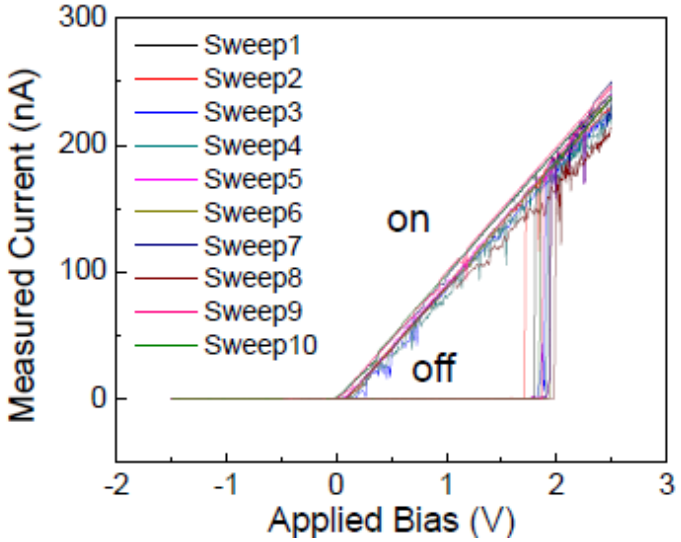
Integrated crossbar/CMOS chip with probe card attached

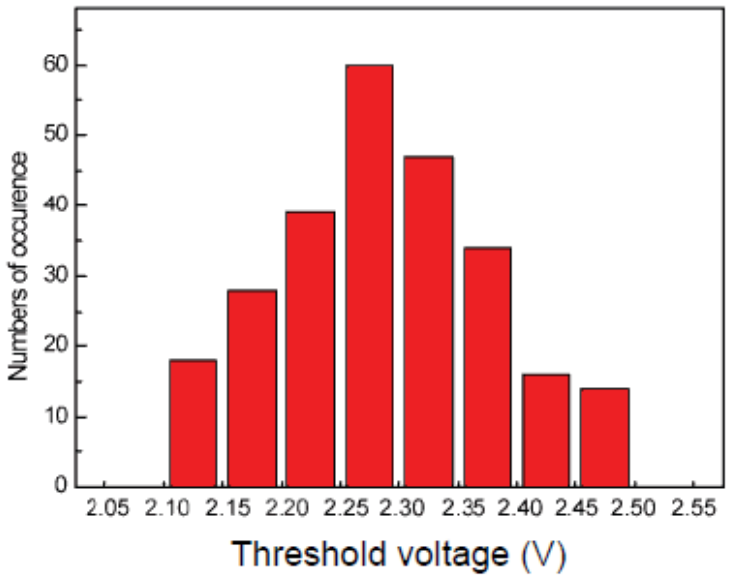
Kim et al. *Nano Lett.*, ***12, 389–395 (2012).***

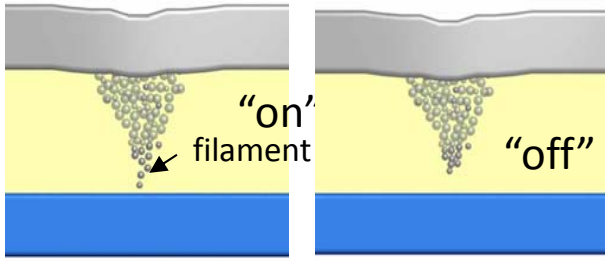
# Integrated Crossbar Array/CMOS System

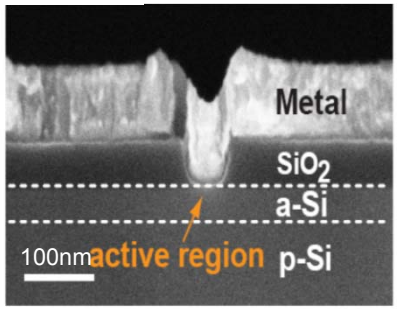


# Performance of a-Si and Metal-Oxide Device Array

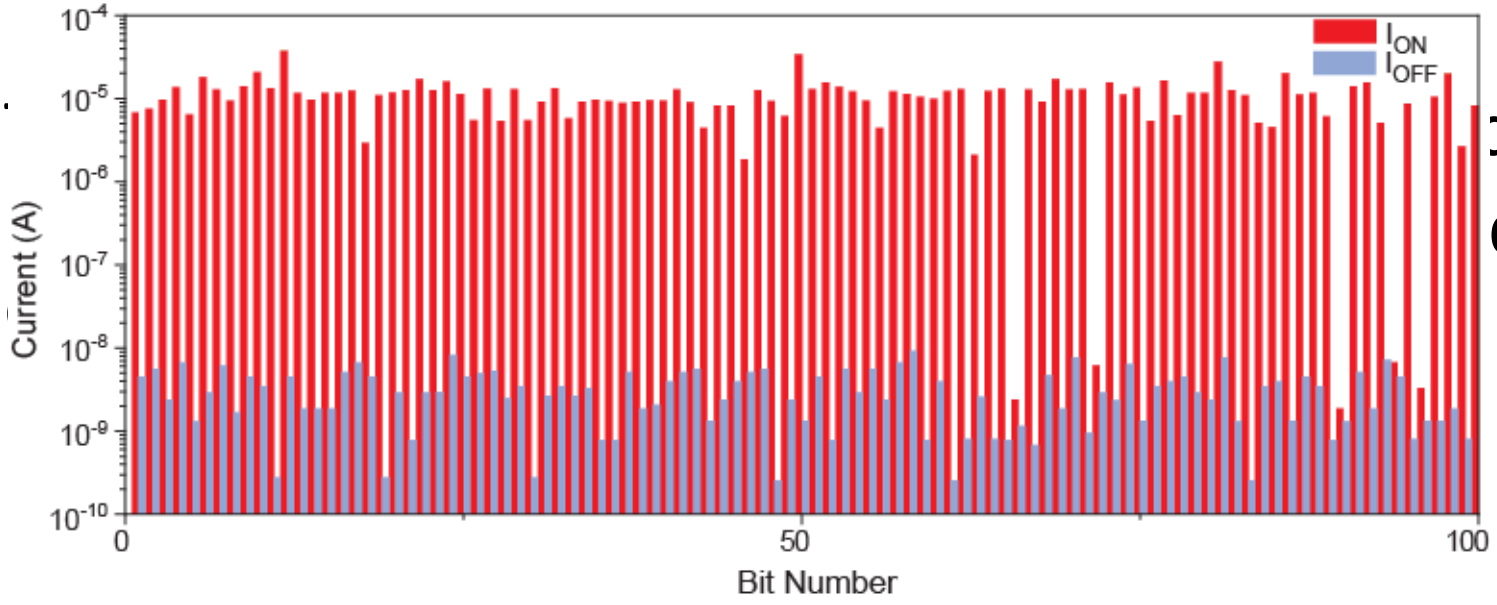


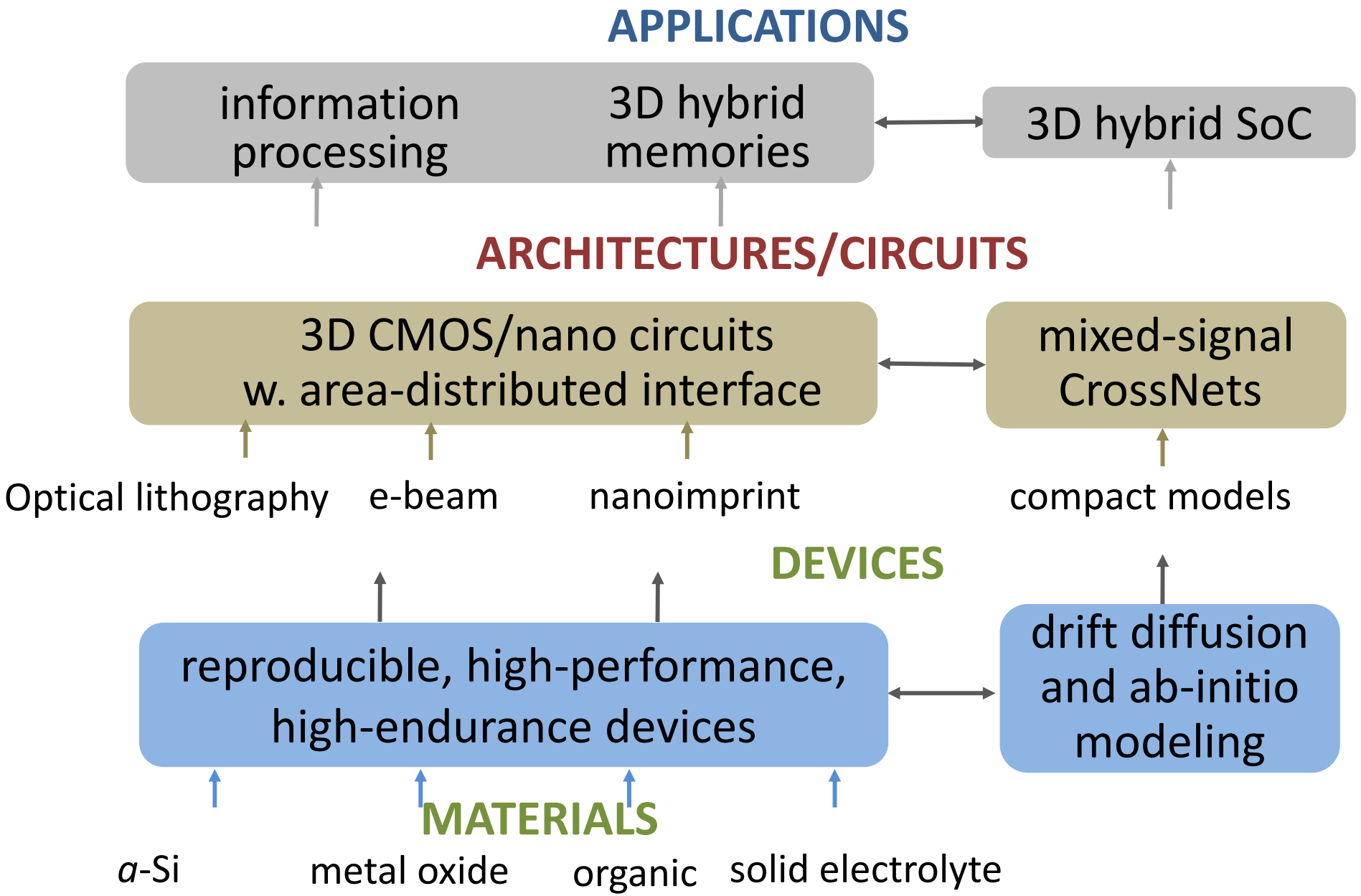




- 
- diode



# Project Overview



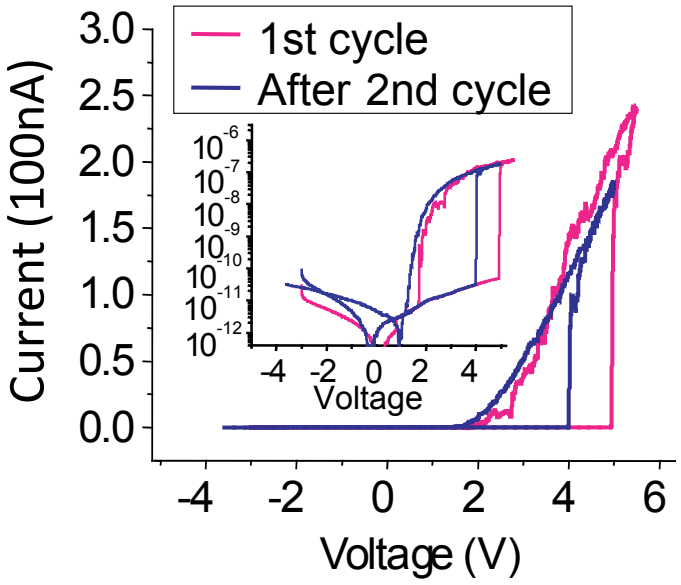
# BACKUP SLIDES

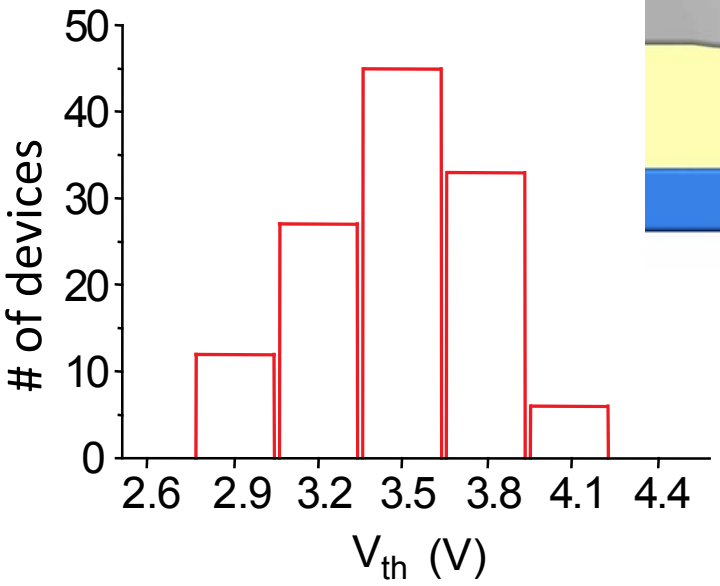
# Thrust Area #3:
## High-Performance/-Yield/-Reproducibility Devices

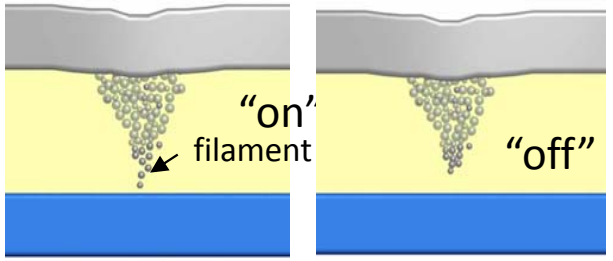
- a-Si (Lu)
- Metal oxide (Stemmer, Xia)
- Organic (Chabinyc)
- Solid-state electrolyte (Lu)

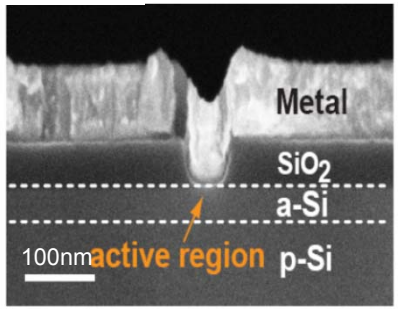
# a-Si Memristive Devices and Arrays

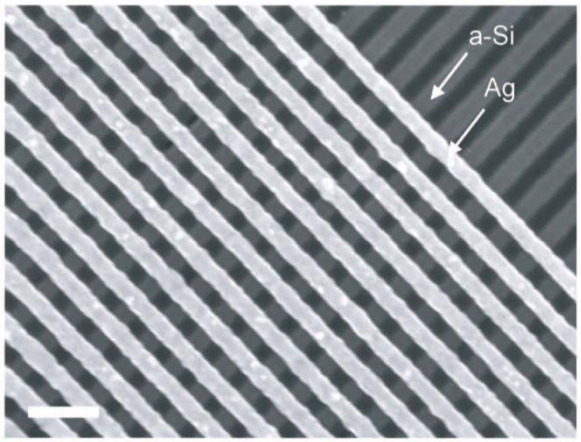
PI: Lu

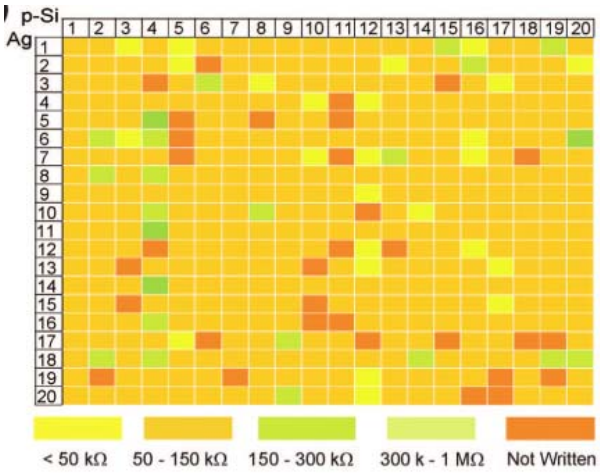



Lu et al., *Nano Lett.* (2008, 2009)

# Project Organization

| UCSB | | | | |
|---|---|---|---|---|
| CMOS circuit design for CMOL integration | MBE fabrication of memristive devices; | Organic memristive devices | Memristive device modeling | Digital and analog 3D hybrid circuit architectures |
| **Cheng, Strukov, Theogarajan** | **Stemmer** | **Chabinyc** | **Strukov** | **Cheng, Strukov, Theogarajan** |

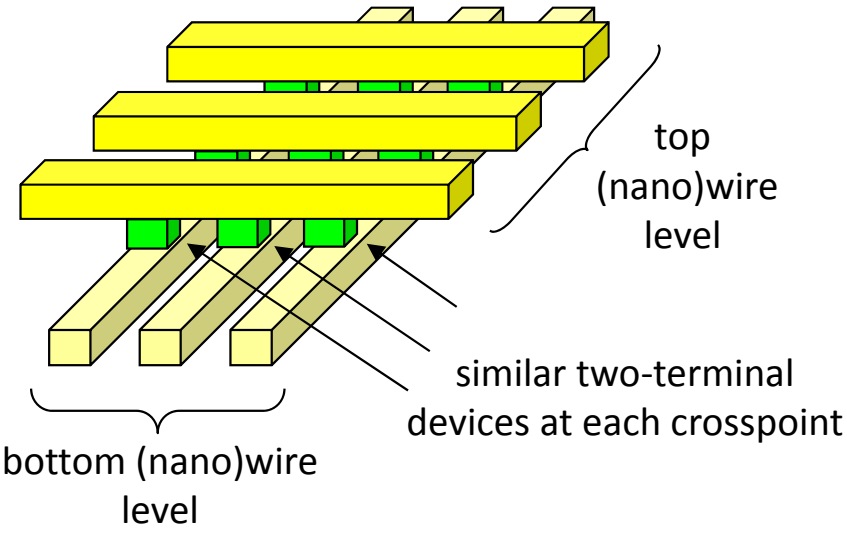
| U. Michigan | UMass | Stony Brook University | |
|---|---|---|---|
| a-Si & solid electrolyte devices; 3D integration with CMOS | Metal oxide memristive devices; 3D integration with CMOS | *Ab-initio* simulation of memristive devices | Mixed-signal neuromorphic 3D hybrid circuit architectures |
| **Lu** | **Xia** | **Fernandez-Serra, Likharev** | **Likharev** |

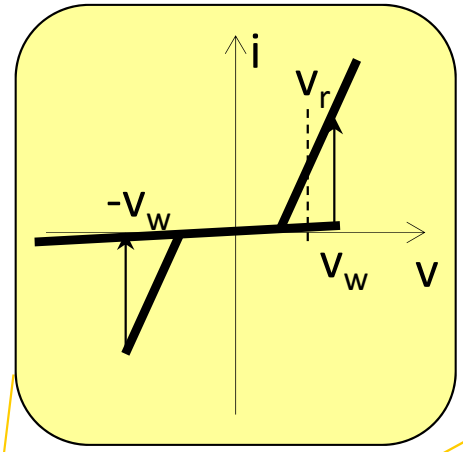


<------- experiment ---------> <---------theory/modeling-------->

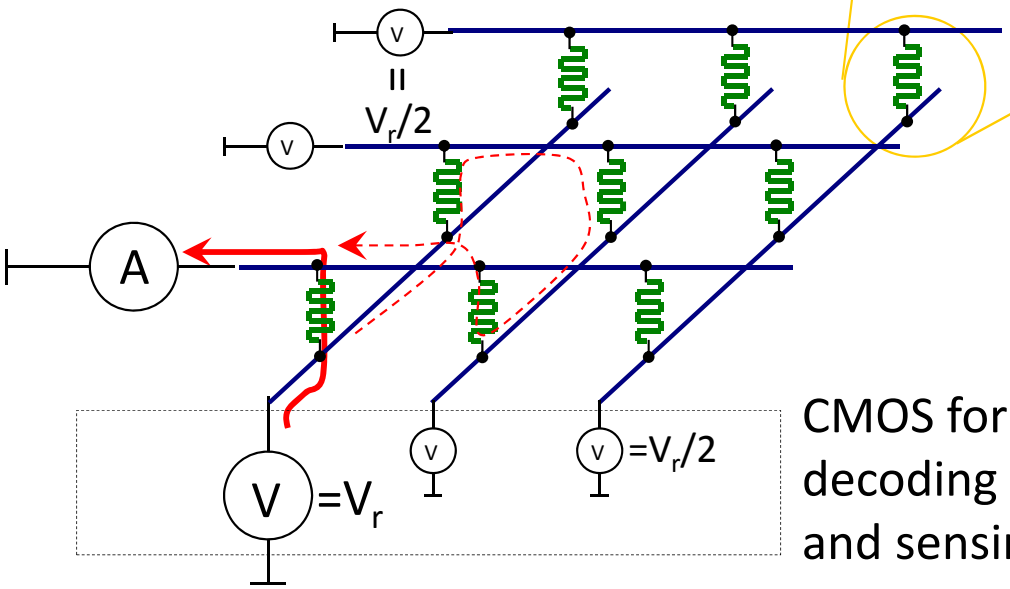
# Crossbar Architecture

- Xbar to preserve density
- Passive (no transistors) but nonlinear I- V
- Common way (from periphery)



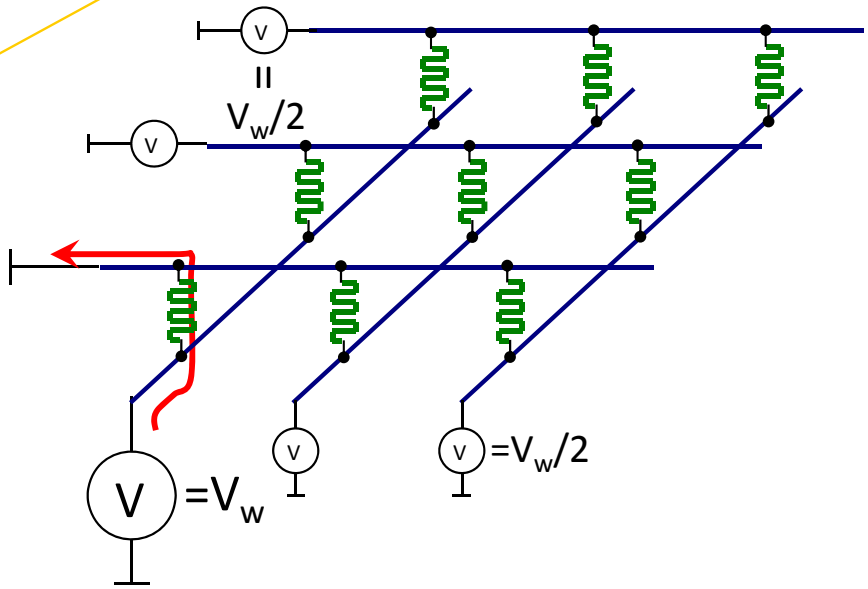




**Read**



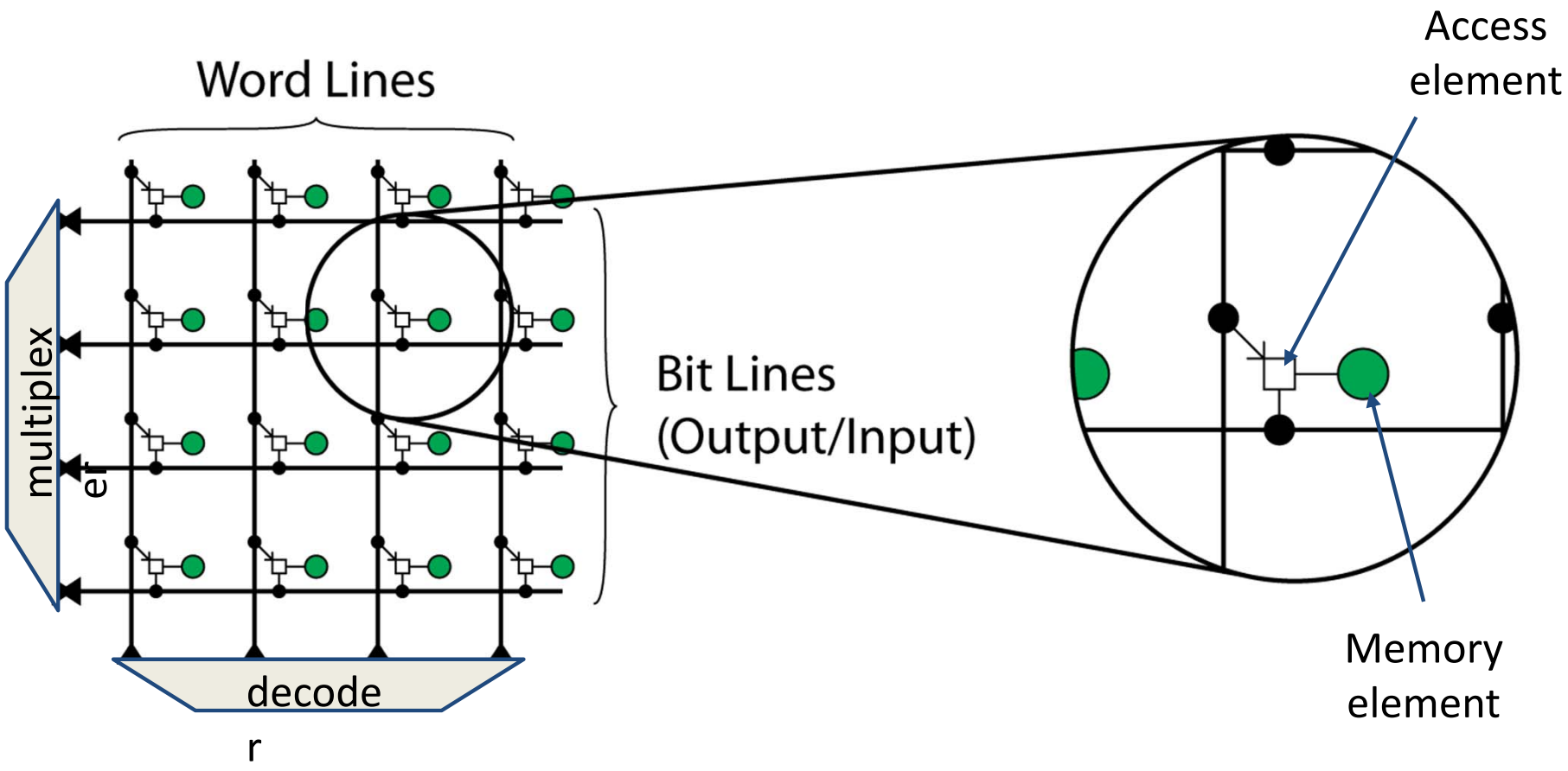

CMOS for decoding and sensing

**Write**



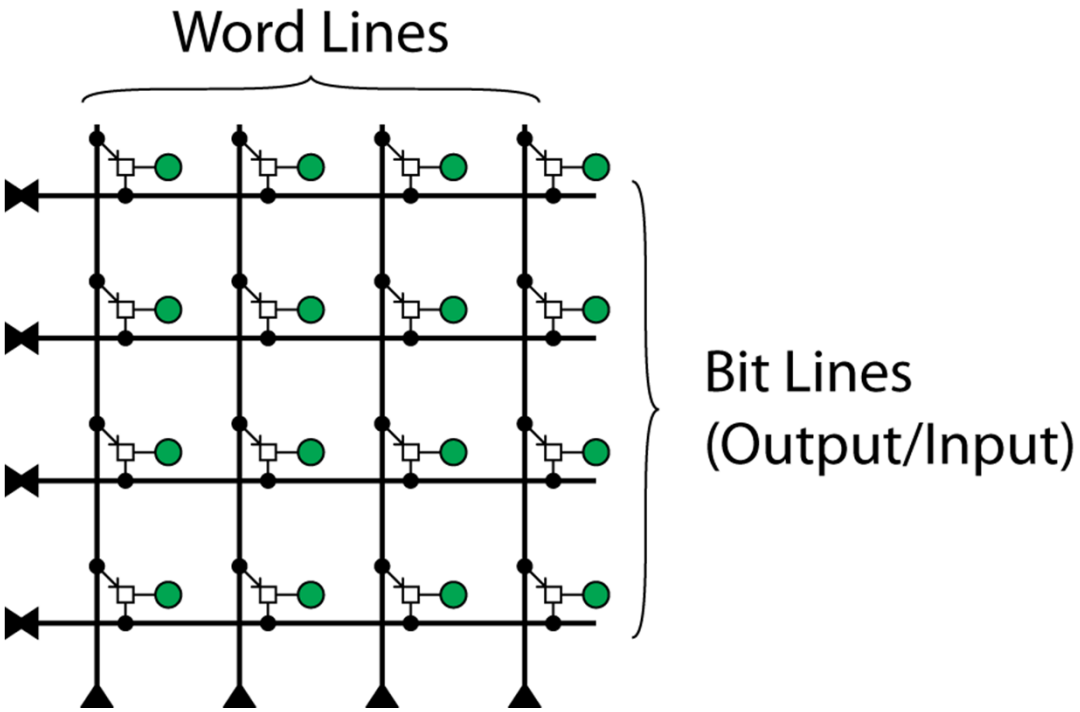
# Generic Memory Array

- Asserting a **word line** makes the **access element** to place the contents of the **memory element** in the **bit line**. A particular bit is then selected with a MUX.

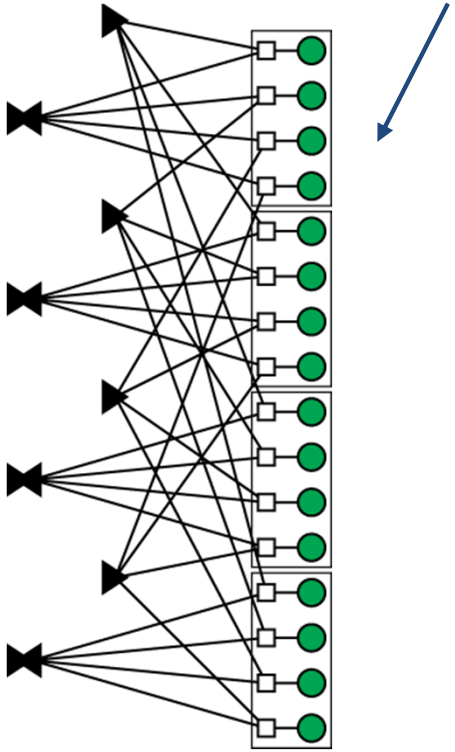


# Generic Memory Array

- An array of $N^2$ memory elements can be **uniquely accessed** using 2N control signals (word+bit lines).

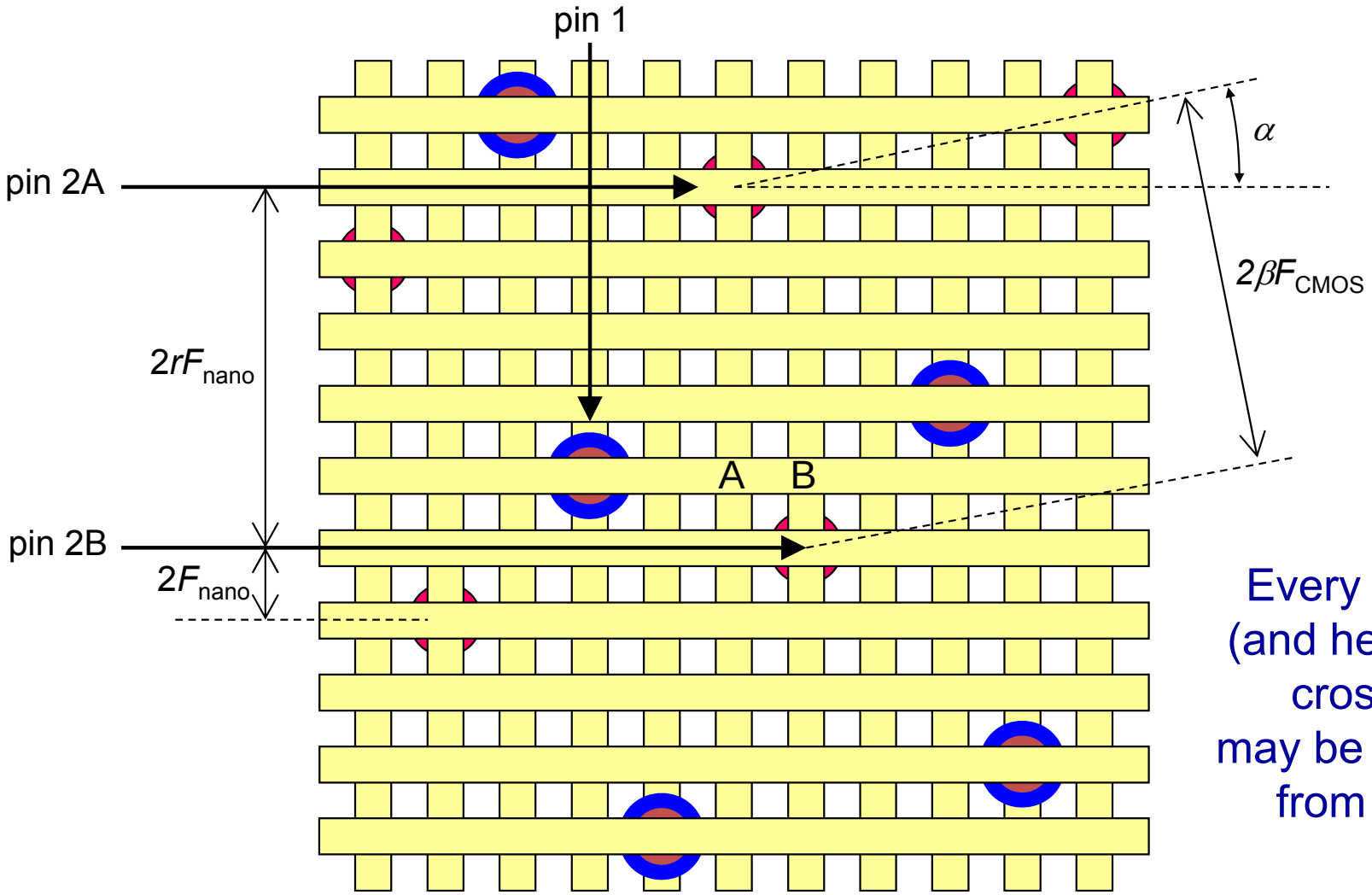



Other representation of the same array



# Area-Distributed "CMOL" Interfaces (II)

Most important feature: pin array tilt by angle $\alpha = \arcsin(F_{nano}/\beta F_{CMOS}) = \arctan(1/r)$



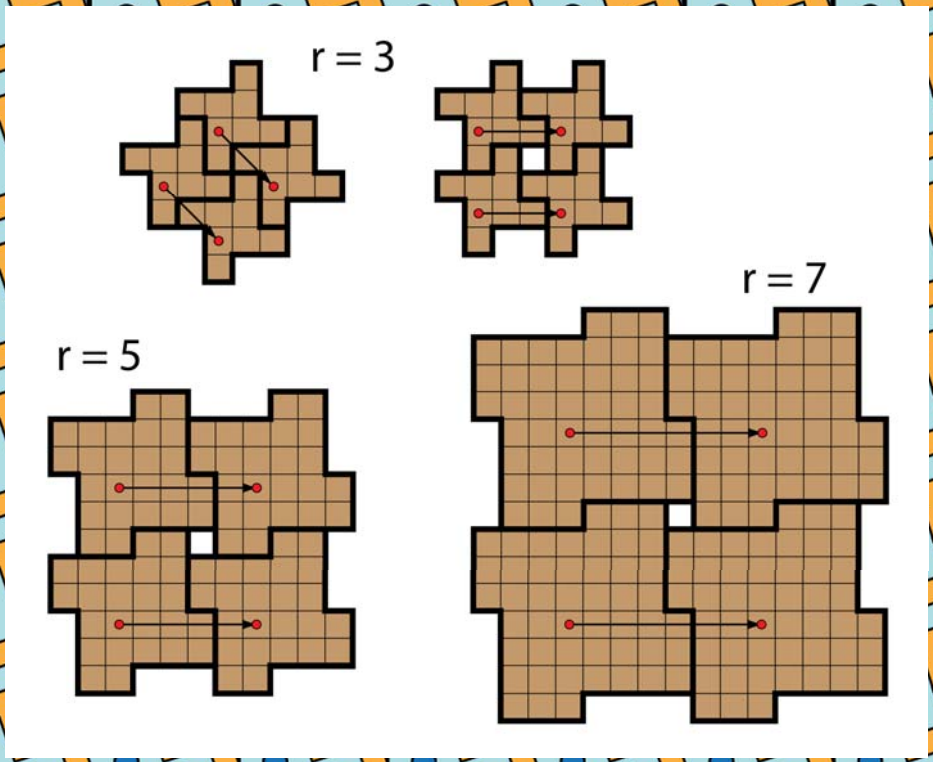

Every nanowire (and hence every crosspoint) may be addressed from CMOS!

K. Likharev (2004, 2005); D. Strukov and K. Likharev (2006)

# A Possible Solution

With this particular connectivity domain geometry (r=3), we can cover all the plane... But that is not always the case.

The pin translation wires are another layer of wires on top of the crossbar




**We can add more crossbar layers by simply inserting a layer of pin translation wires between them.**

# Crossbar Analysis

The crossbar is rotated by an angle α such that:

$$\sin\alpha = \frac{F_{\text{nano}}}{\beta F_{\text{cmos}}} \qquad \cos\alpha = \frac{r F_{\text{nano}}}{\beta F_{\text{cmos}}}$$

Where r is an integer (an odd integer greater than 1).

Once we set r and β (the CMOS cell complexity), the angle α and $F_{nano}$ are fixed as well the length of the wires in the crossbar and the number of memristive devices reachable per wire segment.

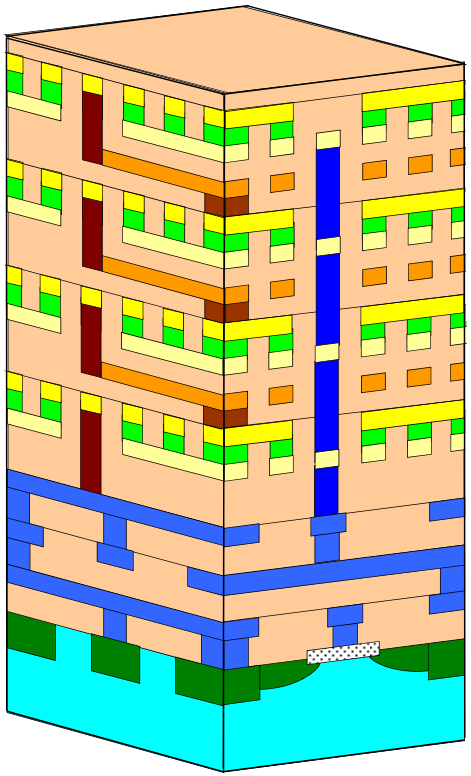
# Crossbar Analysis

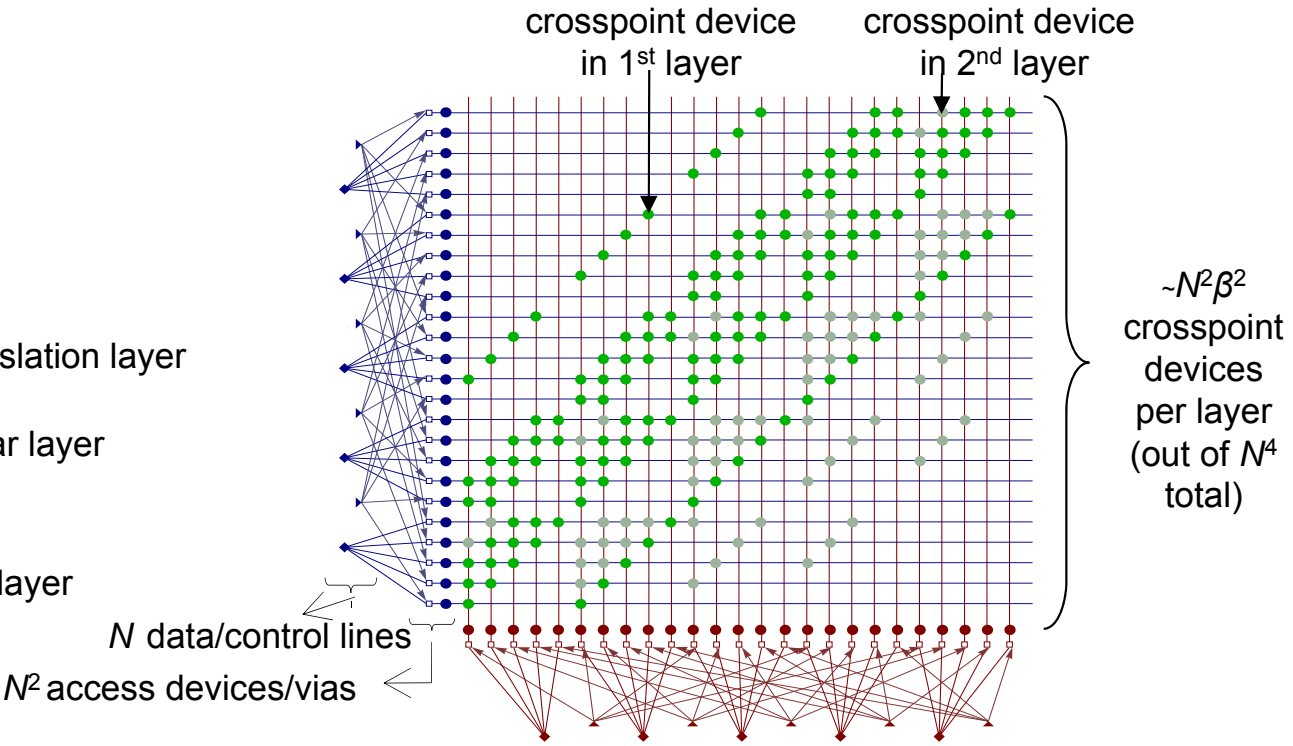
The parameter r also sets the maximum, minimum and average paths the electric signals have to propagate to access a bit (a memristive device).

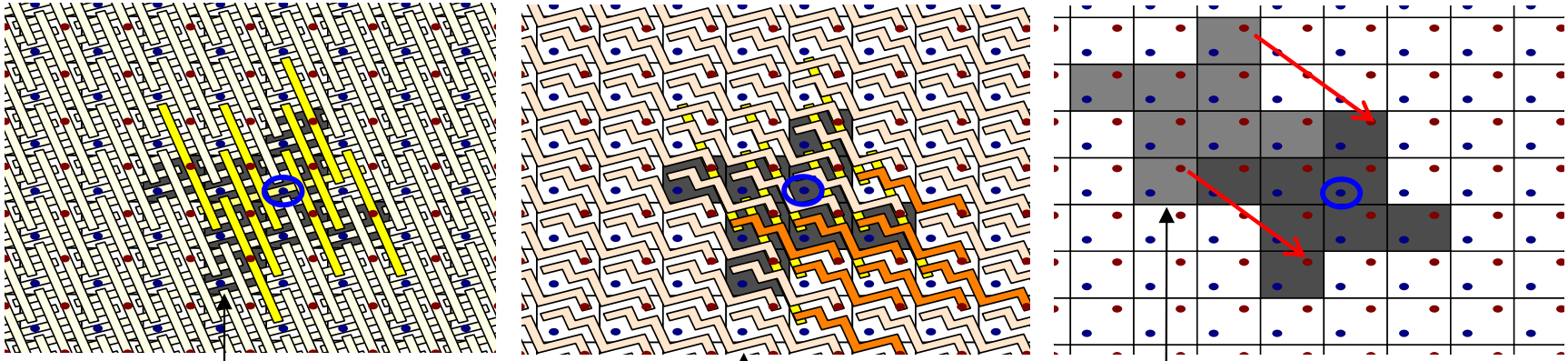
This paths are given by:

Maximum (worst) case: $2F_{nano} * (r^2 - r + 1)$
Minimum (best) case: $2F_{nano} * r$
Average (real) case: $2Fnano * (r^2 + 1)/2$

# 3D Hybrid Integration with Multi-Layer Crossbars

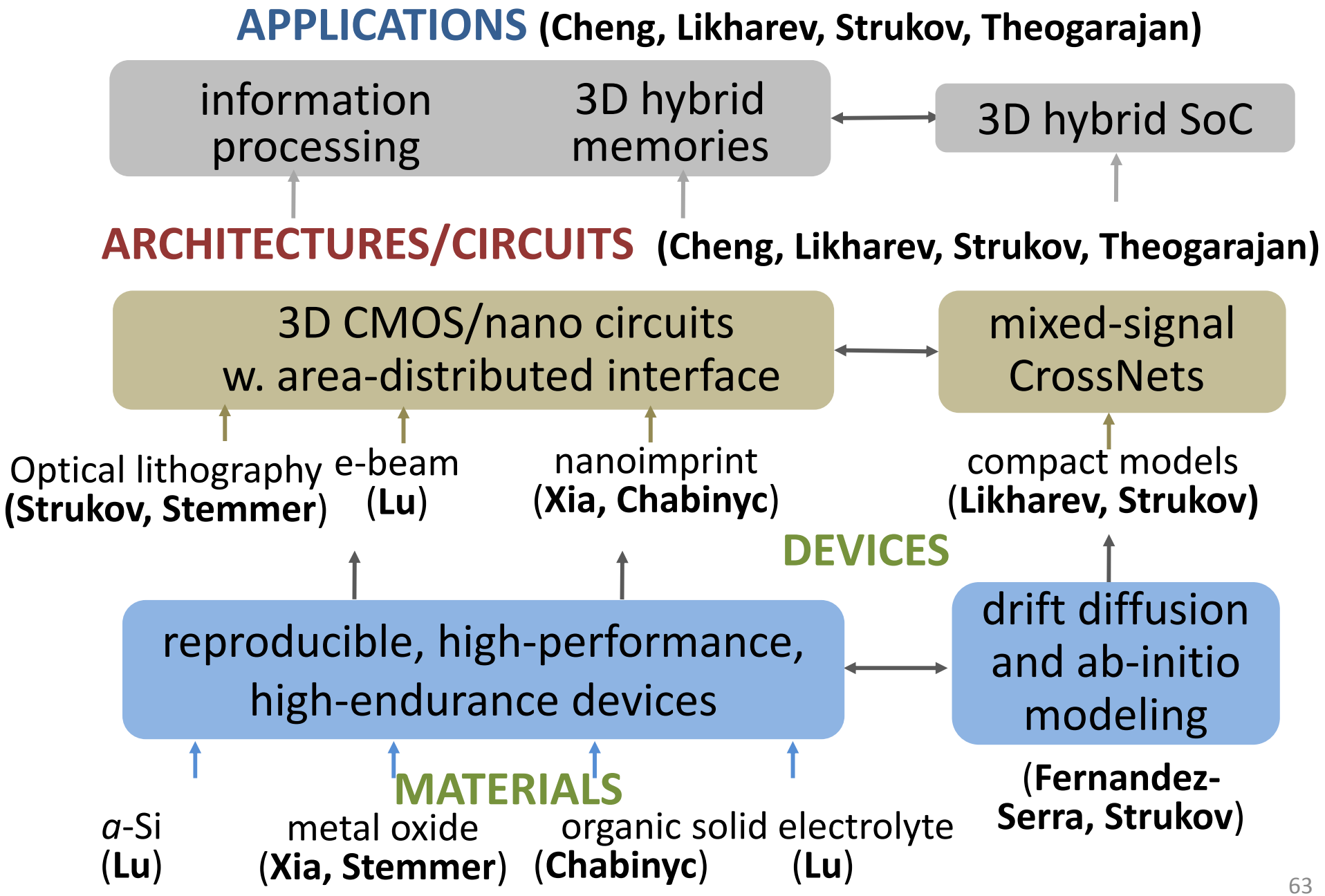


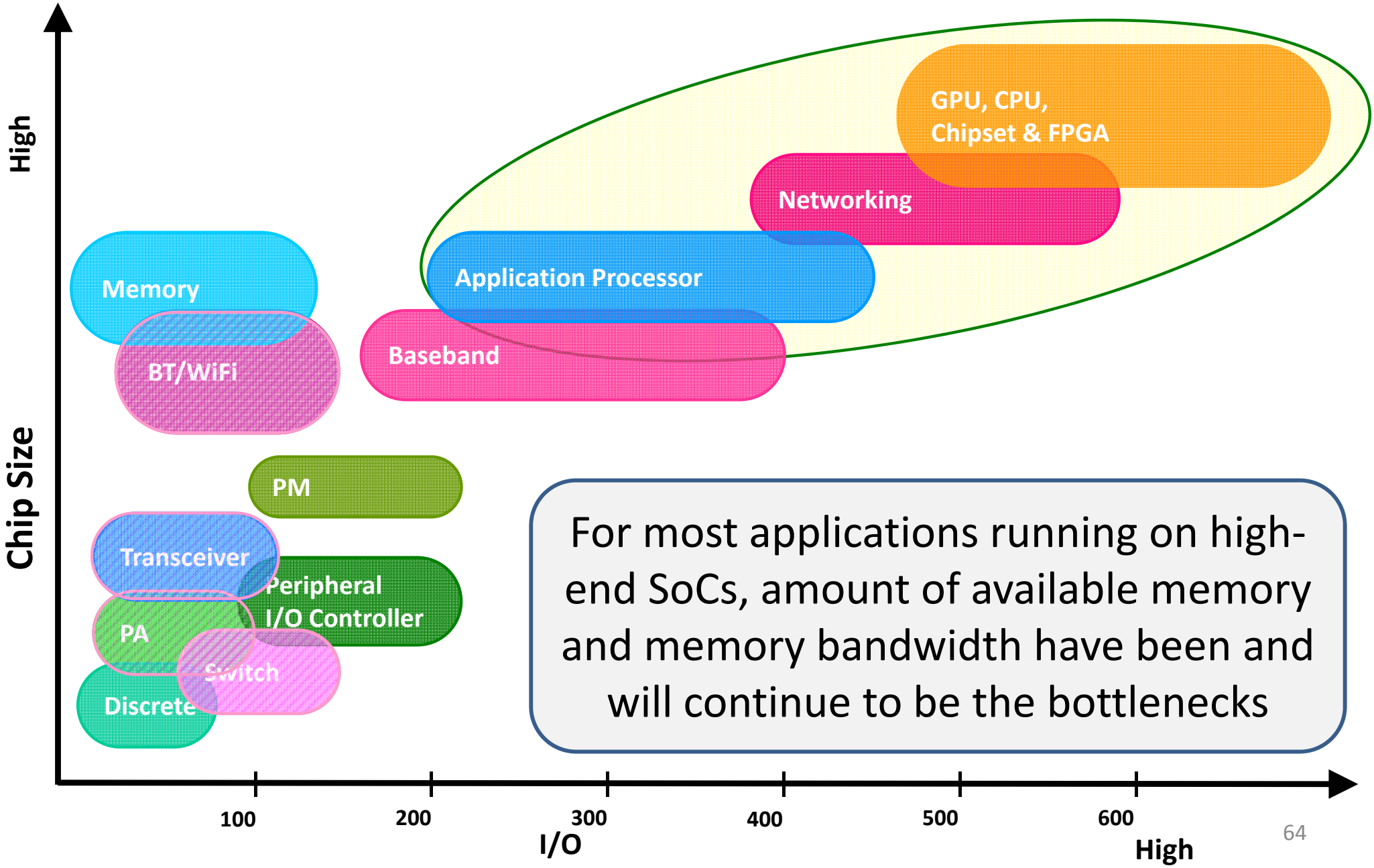




D. Strukov and R. S. Williams (2009)

# Project Overview



# IC Applications Continue to Demand More Memory and Higher Bandwidth



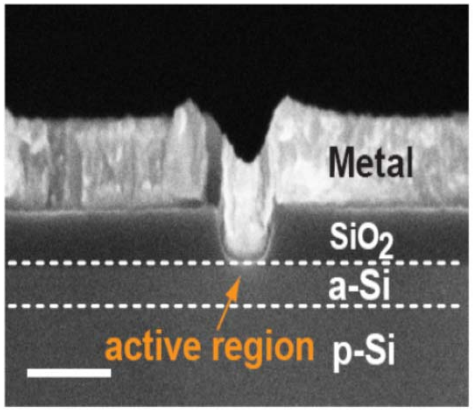

For most applications running on high-end SoCs, amount of available memory and memory bandwidth have been and will continue to be the bottlenecks

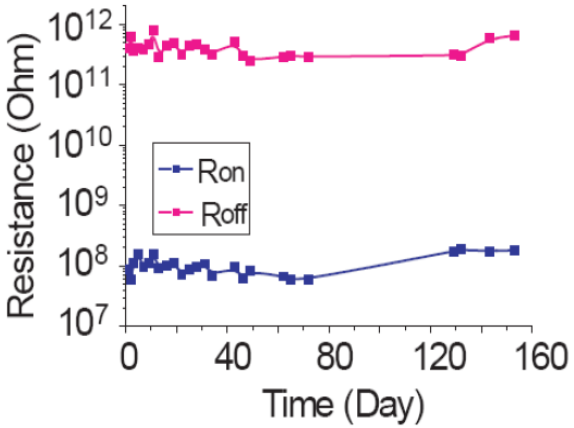
# Bistable Two-Terminal Devices

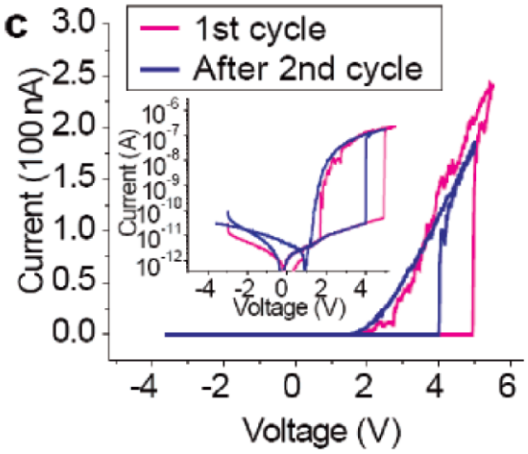
(*latching switches*, a.k.a. *resistive switches*, a.k.a. *programmable diodes, a.k.a. memristive switches*)

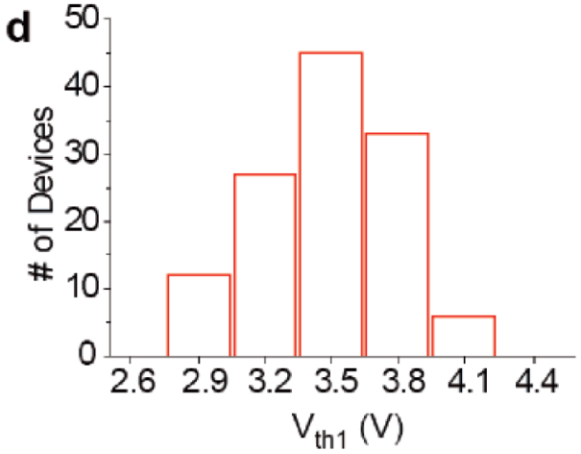
Demonstrated with many materials; no clear winner yet; few reproducibility reports, e.g.:

Si /*α*-Si / M:



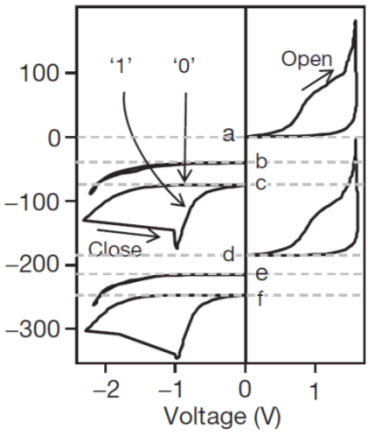



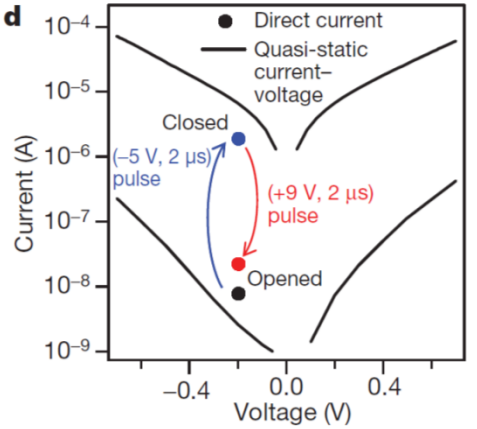



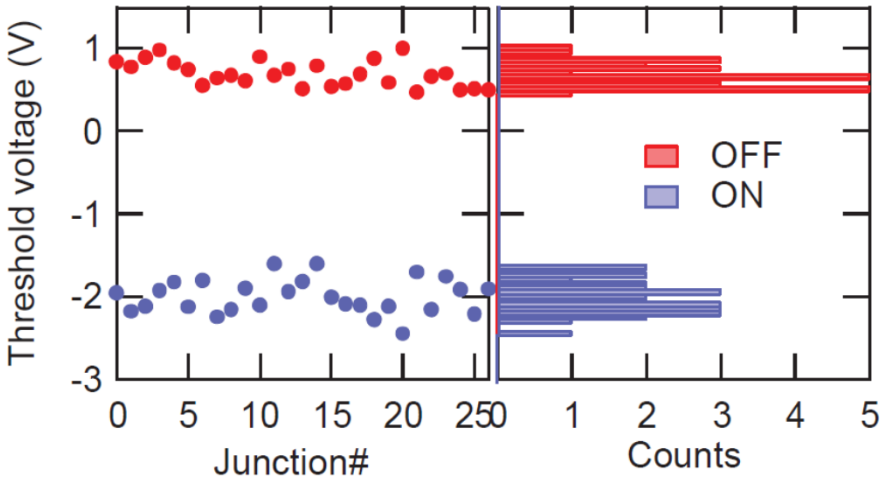




S. H. Jo and W. Lu (2008)

Ti / Pt / $TiO_2$ / Pt:



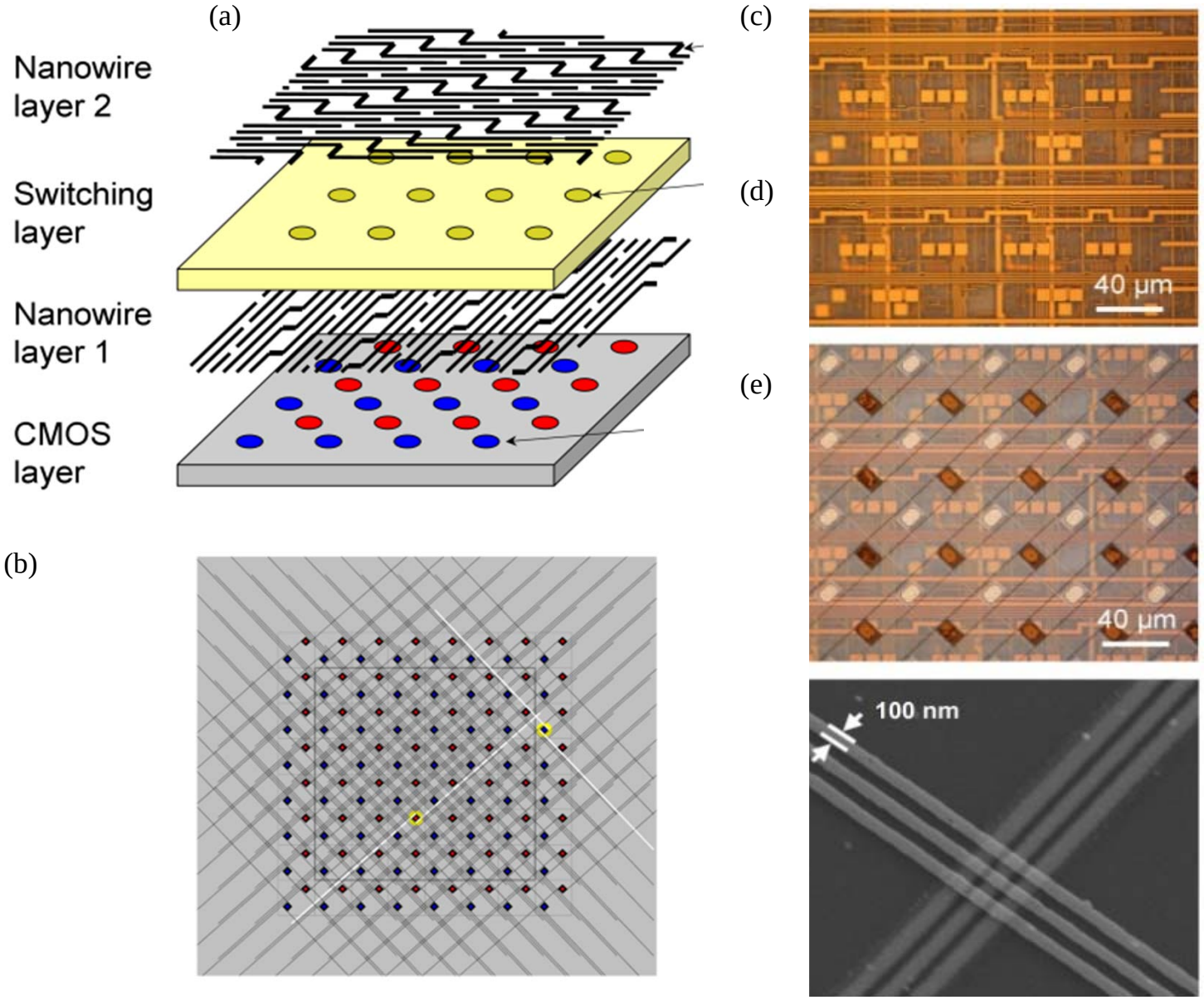

Q. Xia *et al.* (2009);
J. Borghetti *et al.* (2010)

# Device Requirements Vary for Different Ckts/Architectures/Applications

| Dynamic range of resistance | Signal | |
|---|---|---|
| | DC | AC |
| Small | Tuning | - |
| Large | Memory, FPGA, DAC | MAC |

# CMOS-CMOL Integration: Initial Demonstration

PI: Xia




Xia, Strukov et al. (2009)