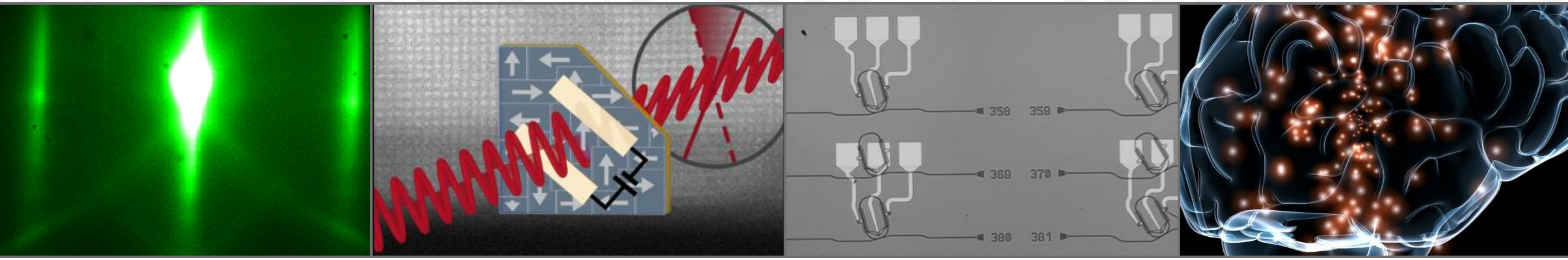
Novel 3D Technologies for the Transition to Cognitive Systems

Bert Jan Offrein



D43D'2017 WORKSHOP

Neuromorphic Devices and Systems Group

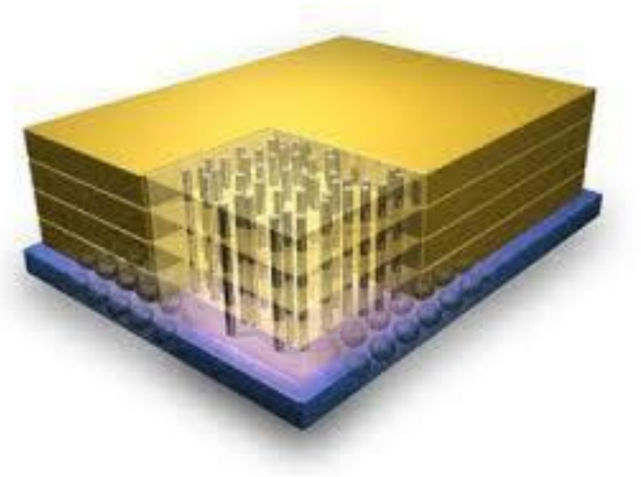
Outline

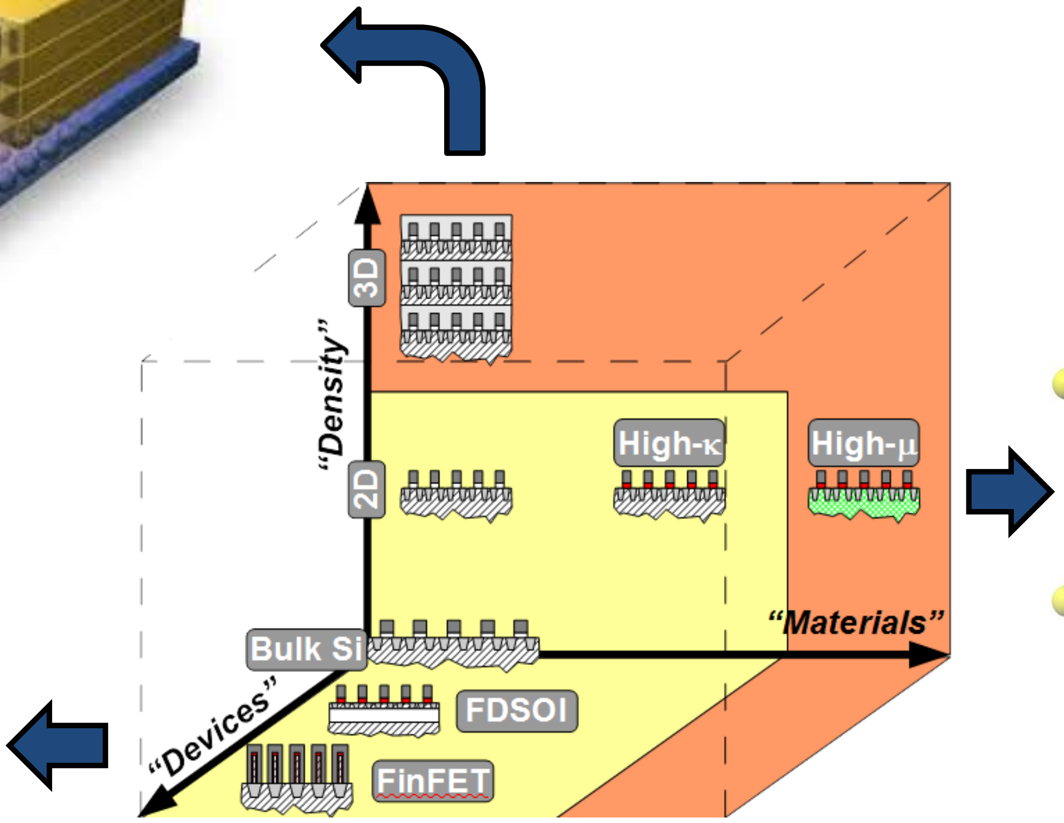
- Chip-level 3D Technologies
  - Introduction
  - 3DM for enhanced CMOS performance

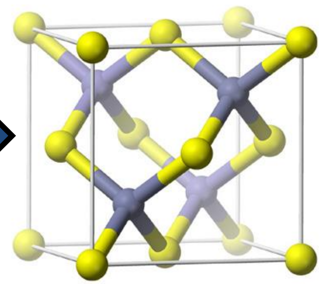
- Enhanced CMOS functionality
  - CMOS Silicon Photonics
    - Monolithic laser integration
  - Novel hardware for Neural Networks
    - Resistive Crossbar Arrays
    - Photonic Synaptic Processor

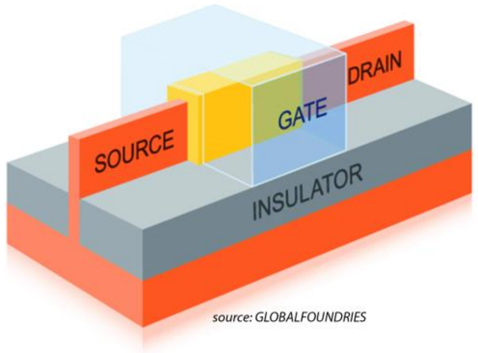
- Conclusions

What would be the ultimate CMOS Technology?



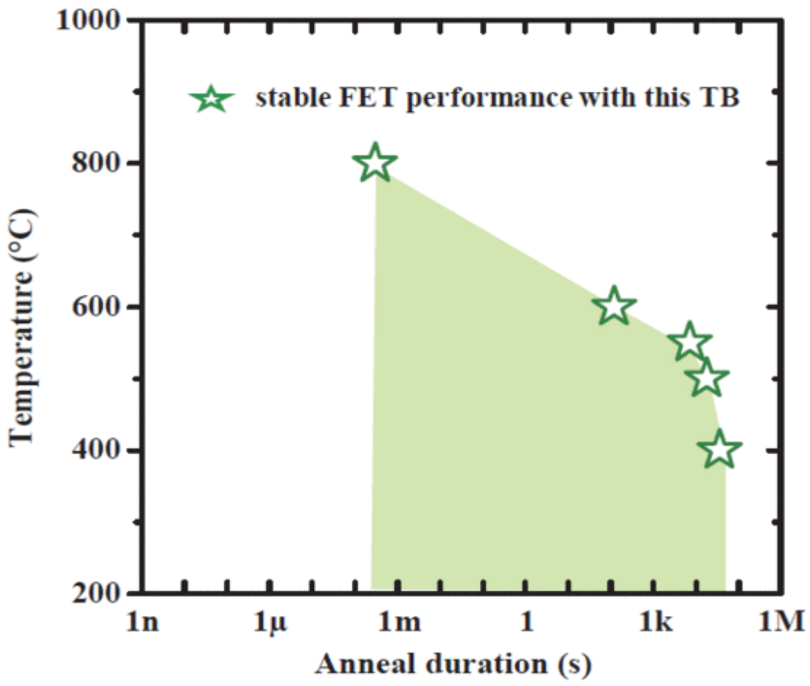






Challenge for 3D Monolithic (3DM)

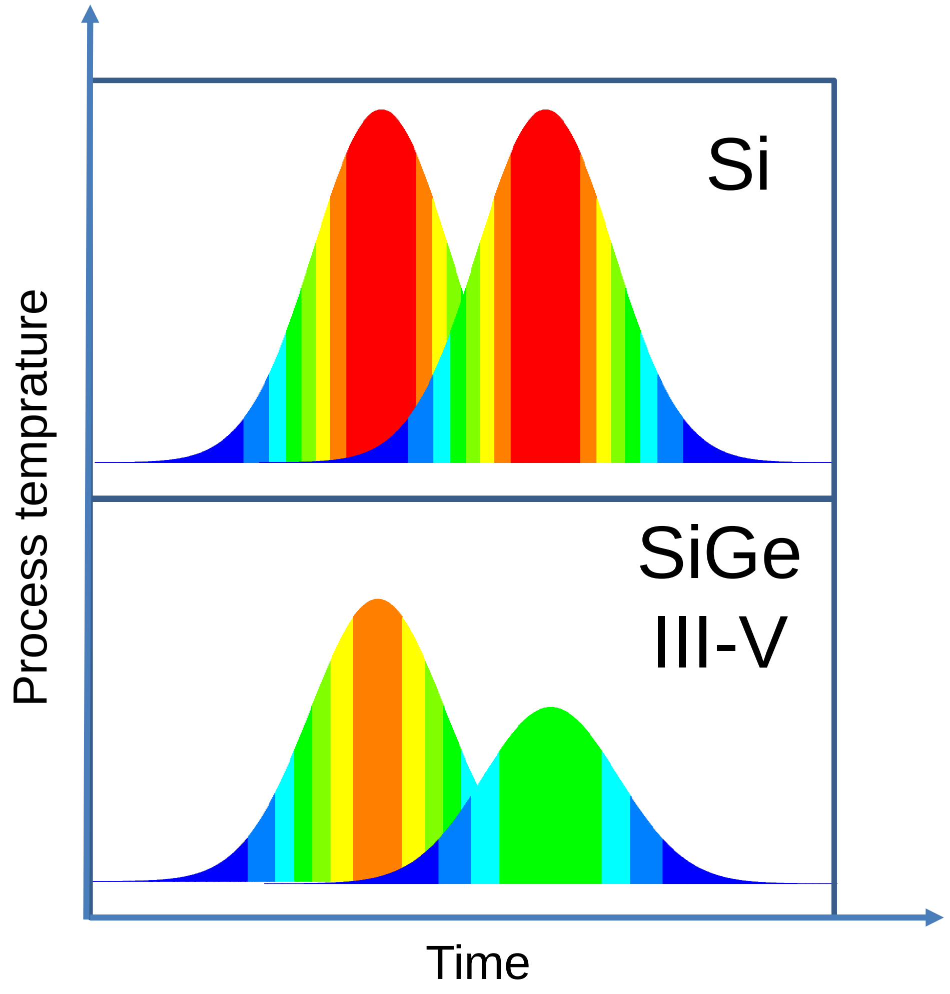
Thermal budget stability of bottom FETs

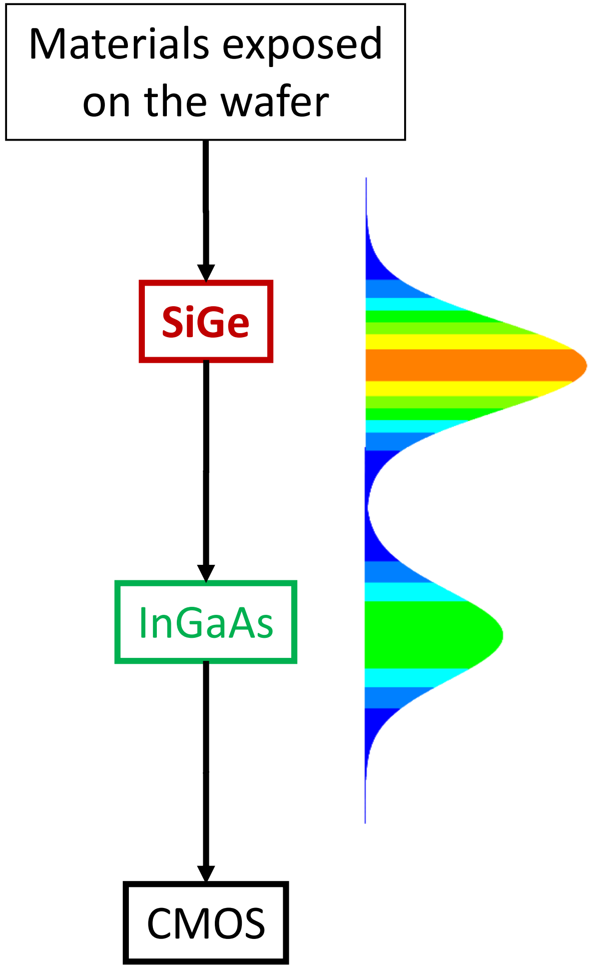


- 500°C for 5 hours
- 550°C for 2 hours
- 600°C for 2 minutes

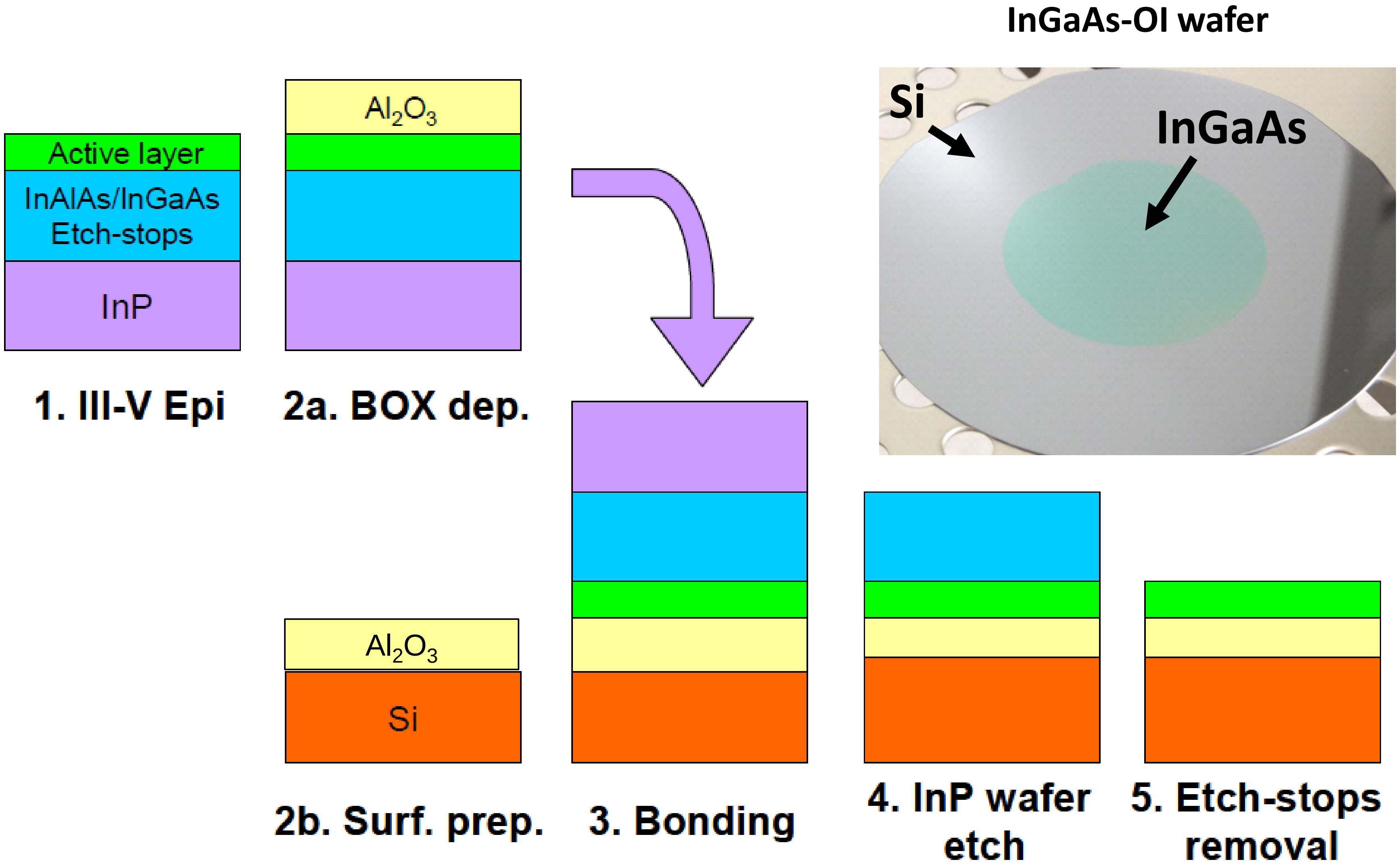
P. Batude, *et al.*, VLSI 2015

Temperature flow – bottom to top



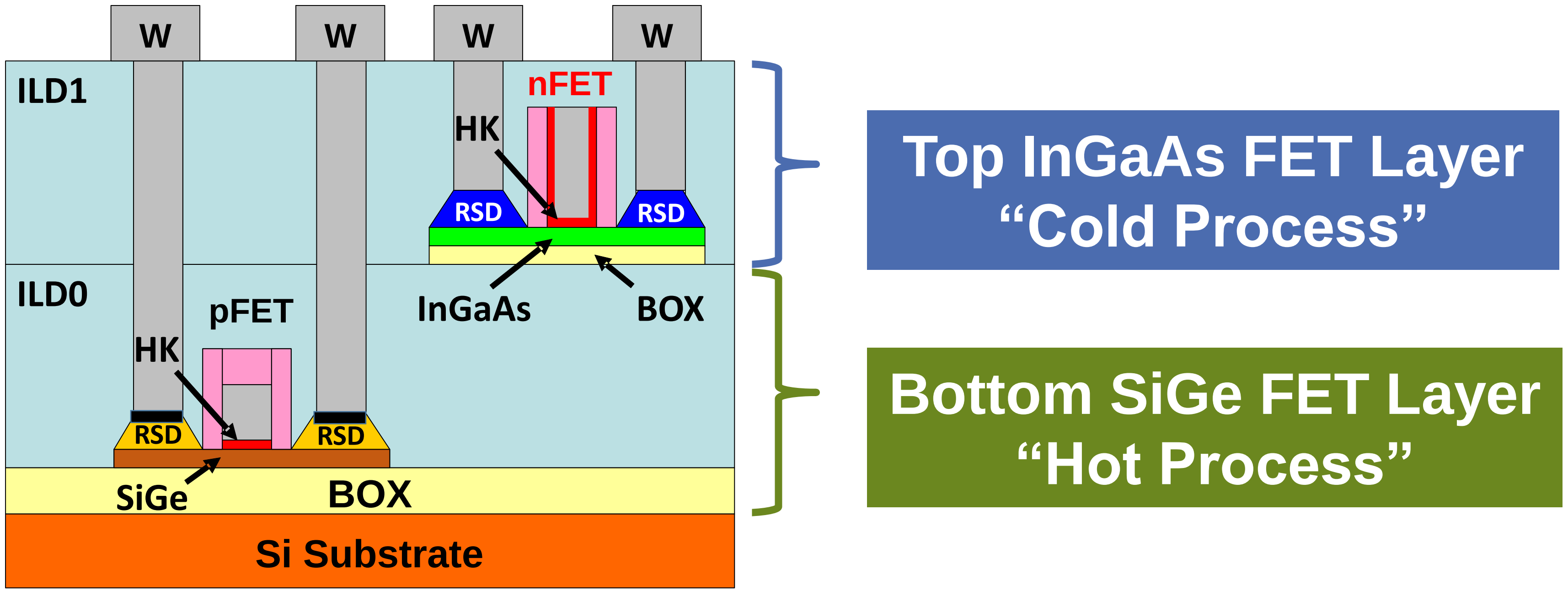


Direct Wafer Bonding



[1] M. Yokohama, et al., VLSI 2010, [2] L. Czornomaz, *et al.*, ESSDERC 2013

InGaAs-on-SiGe 3DM Integration



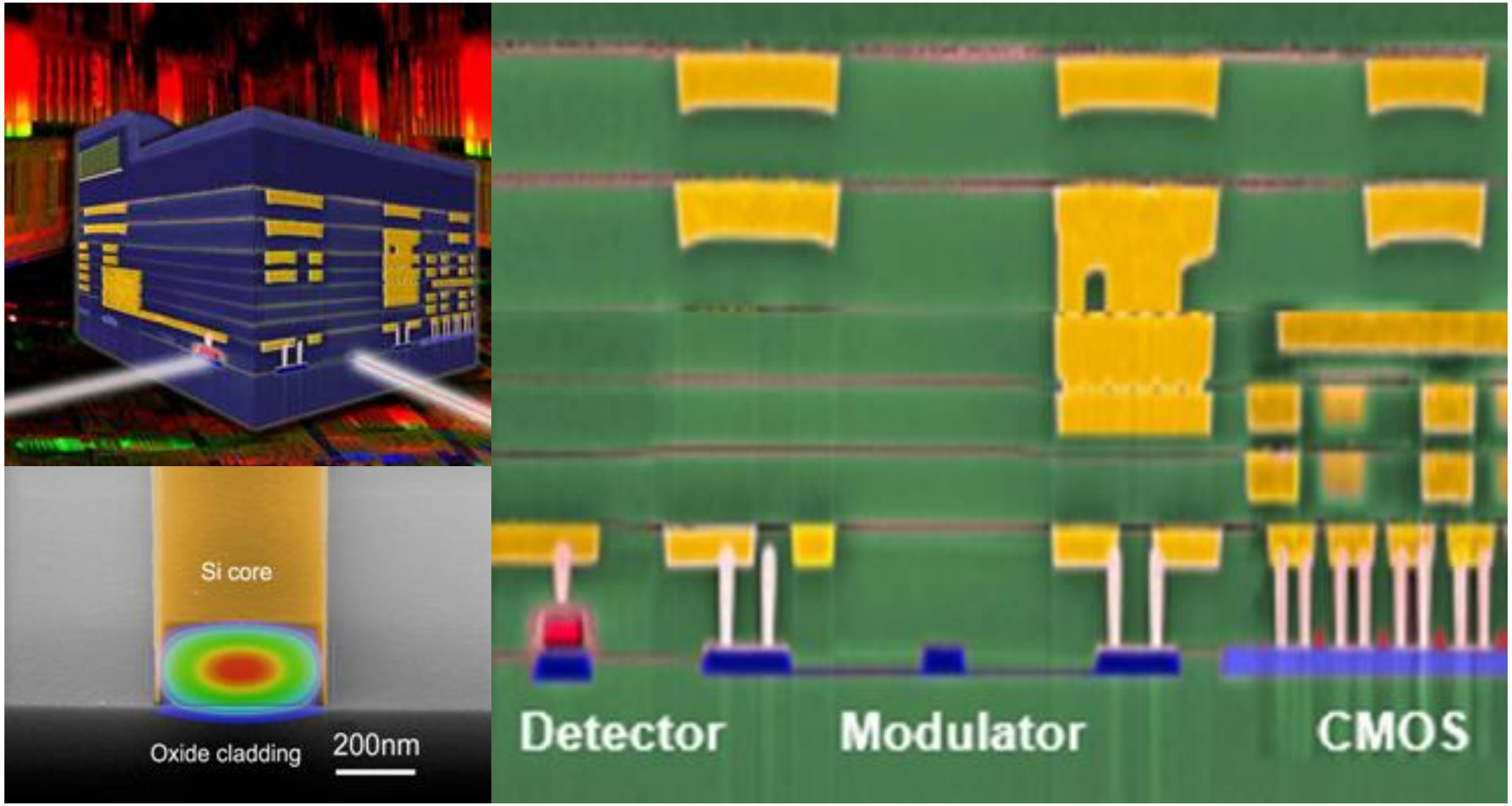
Combined benefits of:

- 3DM integration
- High-mobility channel materials in both layers

Outline

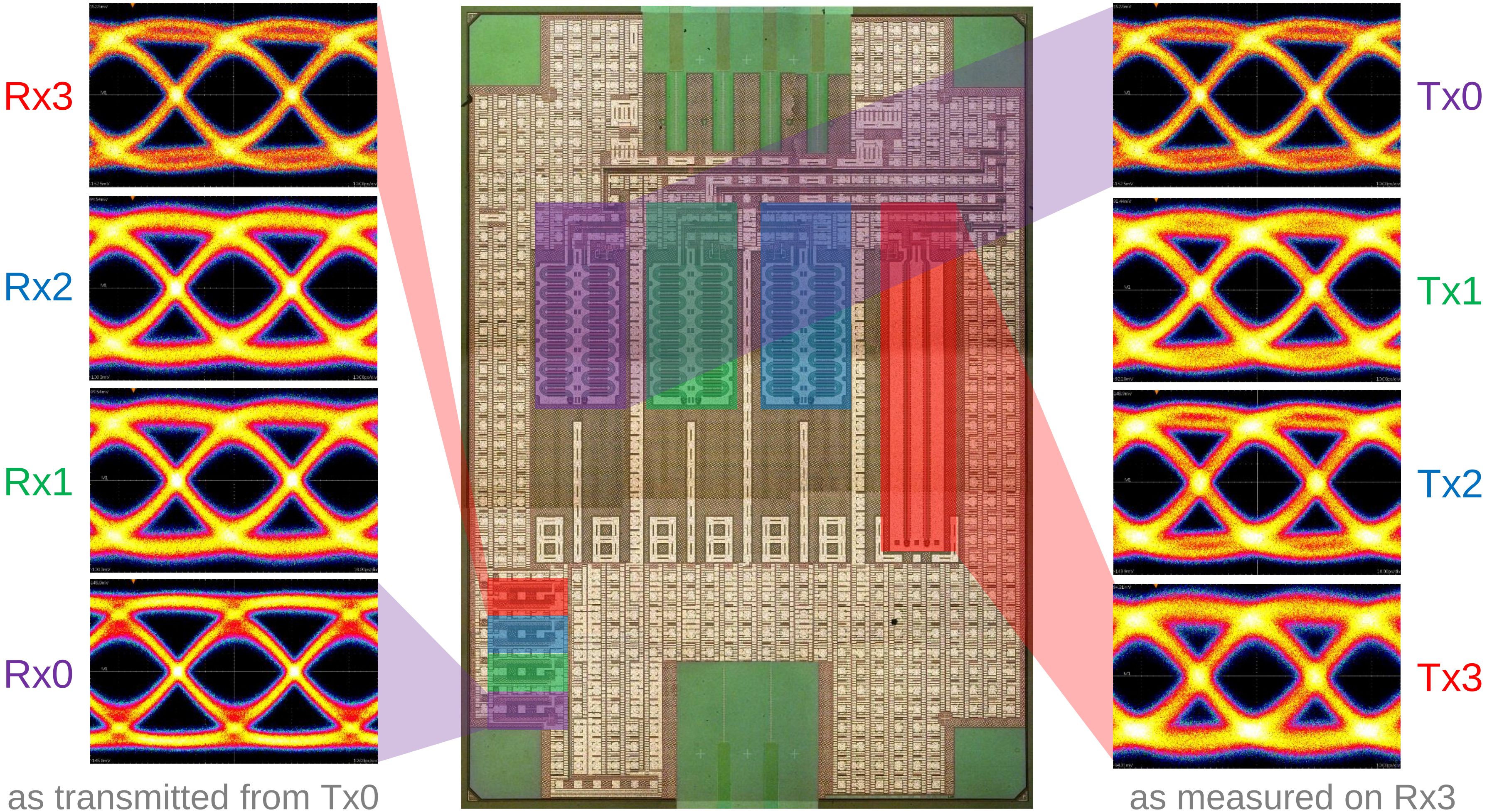
- **Chip-level 3D Technologies**
  - Introduction
  - 3DM for enhanced CMOS performance
- **Enhanced CMOS functionality**
  - CMOS Silicon Photonics
    - Monolithic laser integration
  - Novel hardware for Neural Networks
    - Resistive Crossbar Arrays
    - Photonic Synaptic Processor
- **Conclusions**

CMOS Silicon photonics



Integrate electrical **& optical** functions in silicon

4 λ x 25 Gb/s optical transceiver demonstration



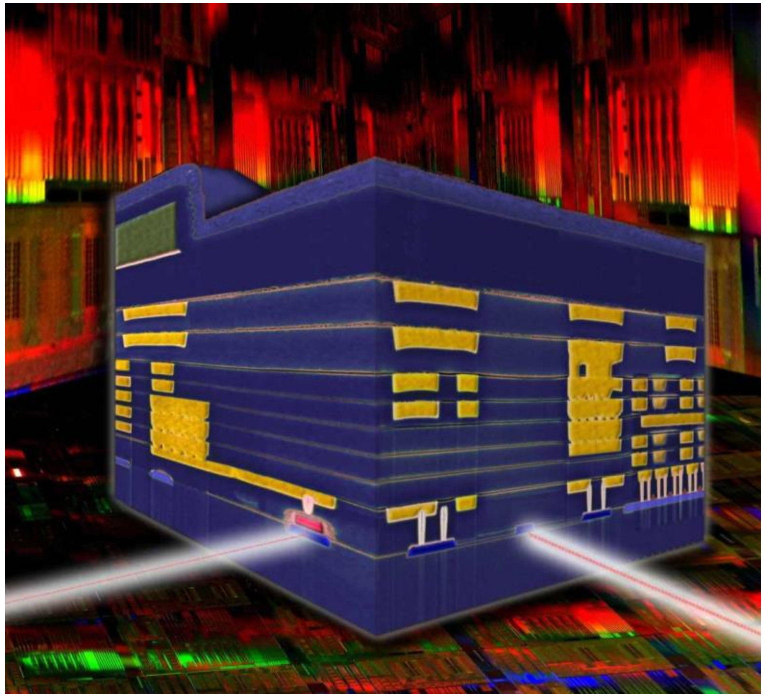
as transmitted from Tx0

as measured on Rx3

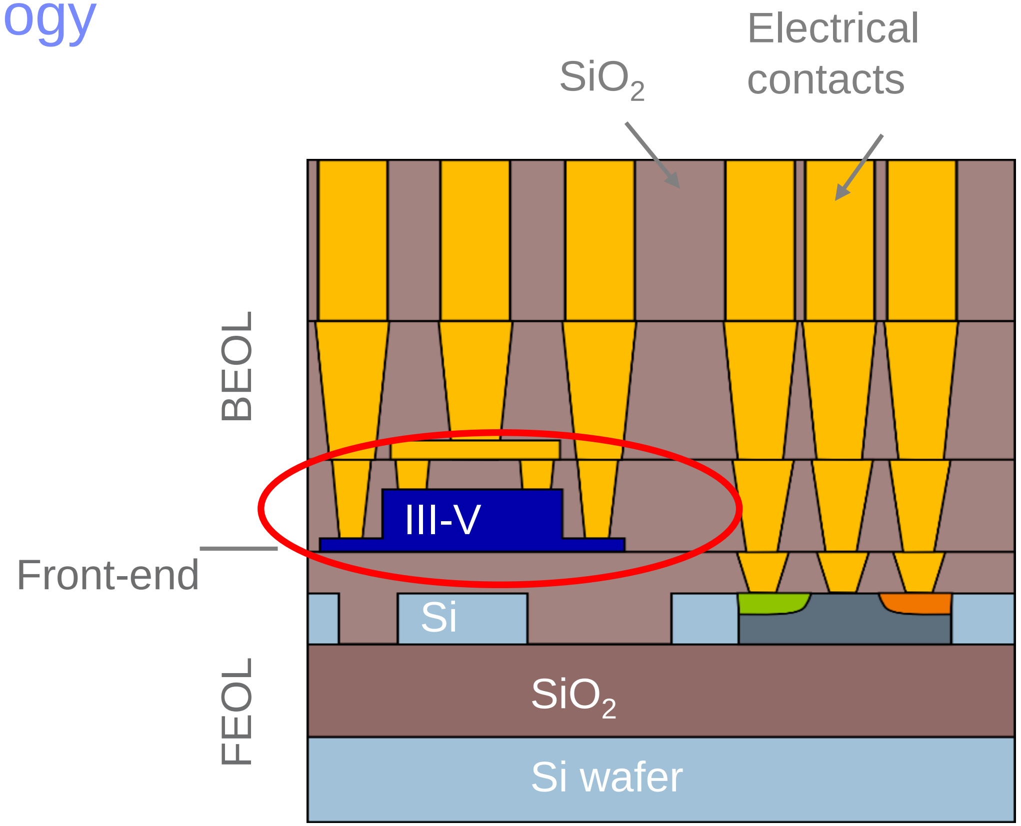
Demonstration of a flip chip mounted 100G transceiver with four wavelength multiplexing at 25 G each.

Bert Jan Offrein

CMOS Embedded III-V on silicon technology



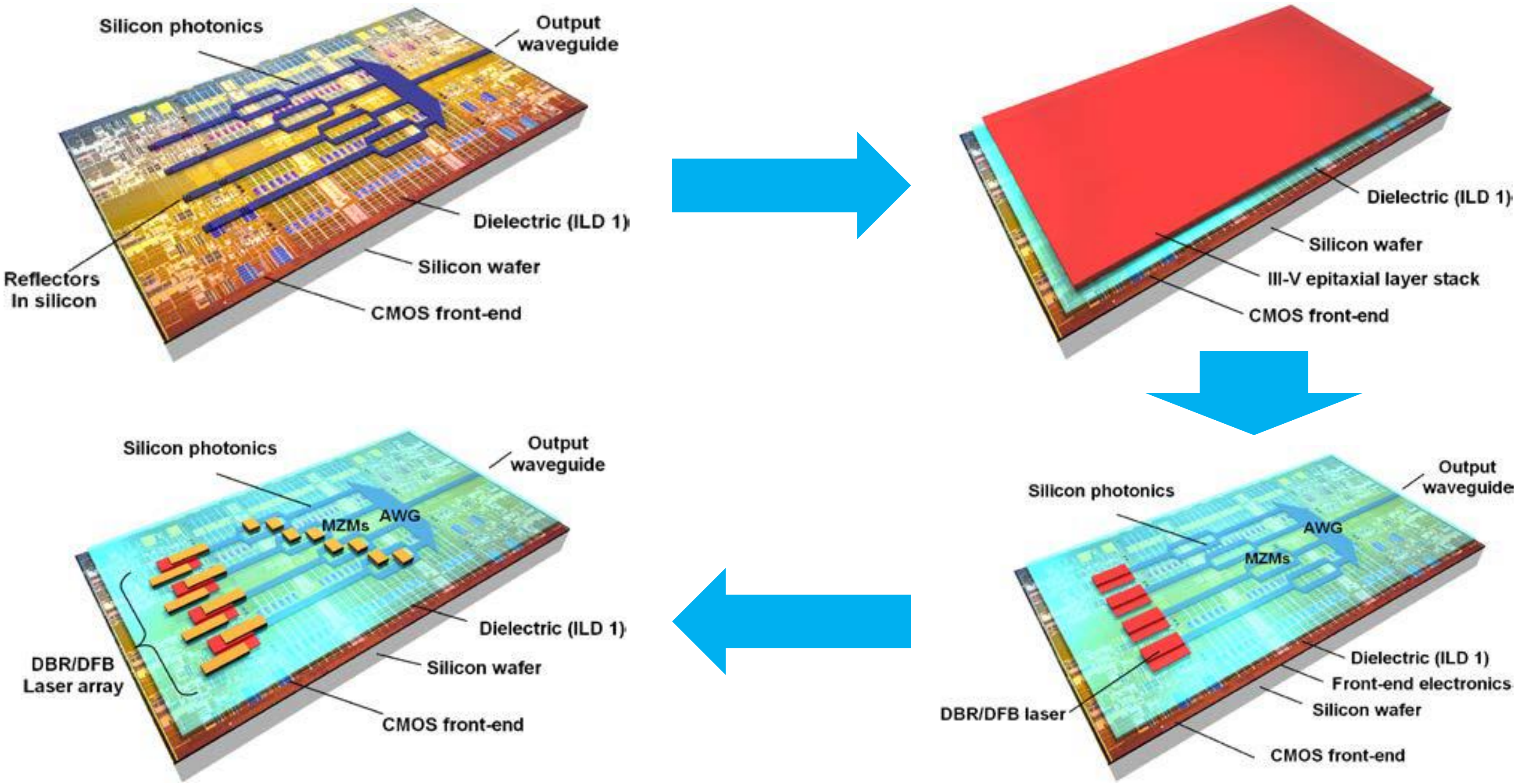
CMOS Si Photonics



… + III-V functionality

- Overcome discrete laser and assembly cost
- New functions, directly combining electronics, passive and active photonics

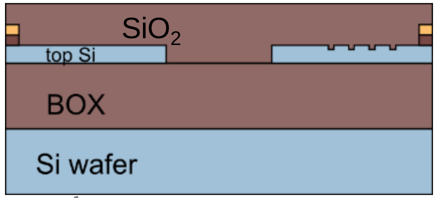
H2020 EU project DIMENSION



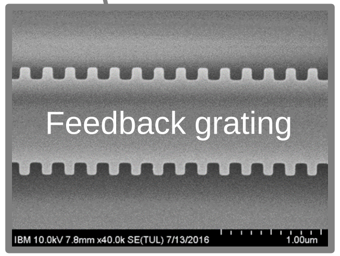
Bert Jan Offrein

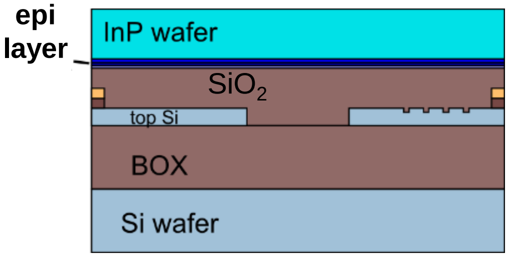
Processing scheme

- 5 InAlGaAs quantum wells (MOCVD)

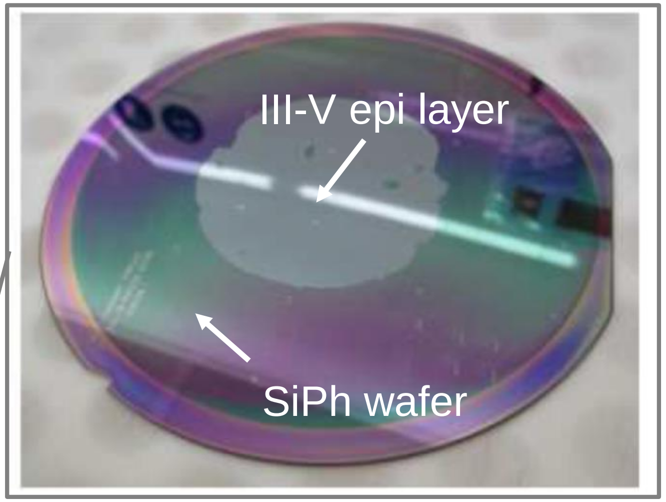


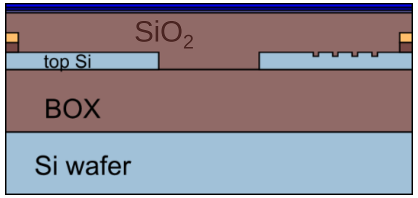
SiPh wafer



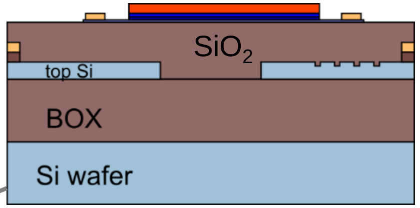


Wafer bonding

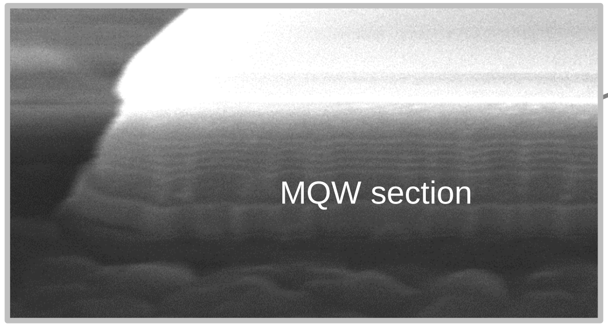


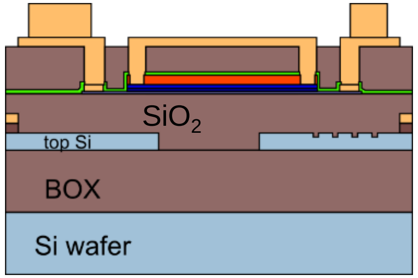


Substrate removal



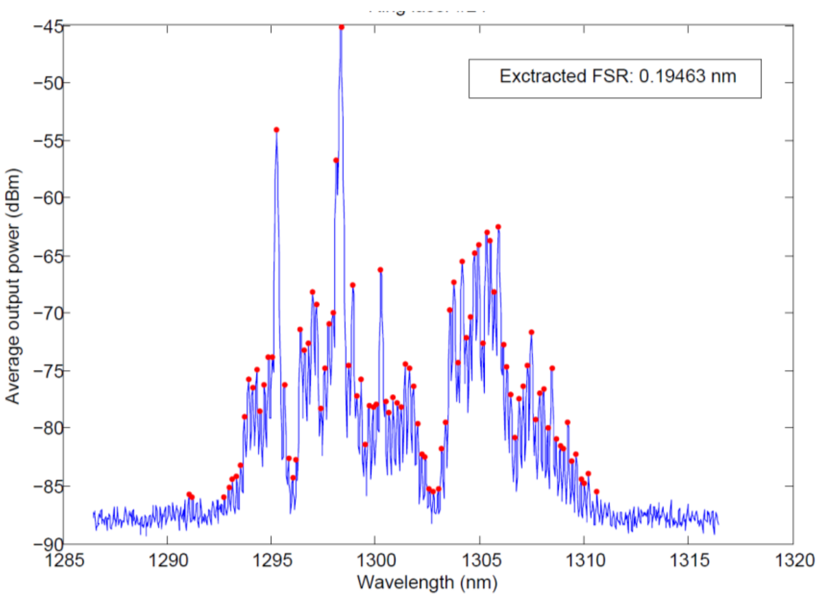
III-V structuring





Metallization

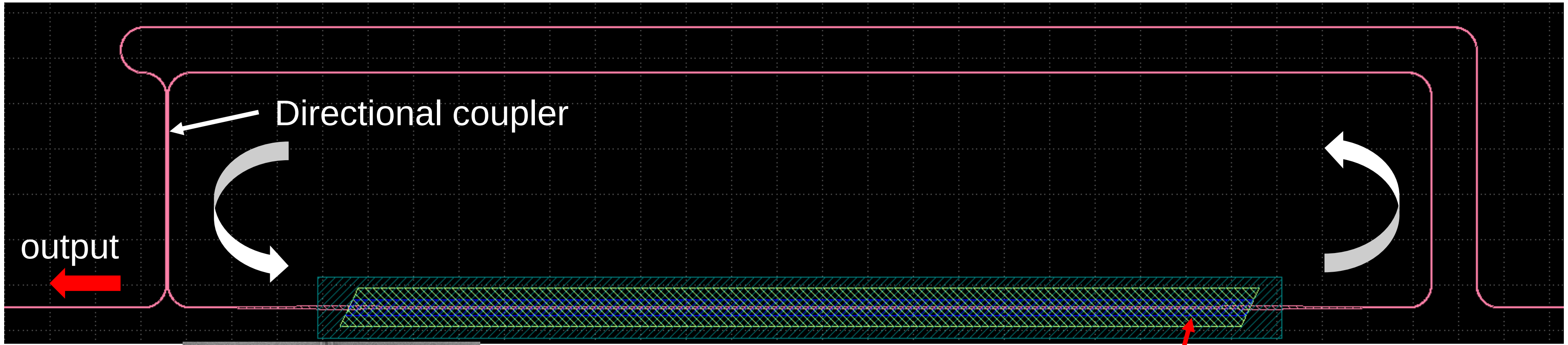
Optically pumped ring laser

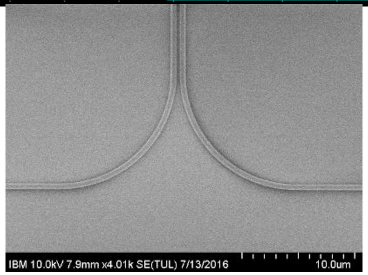


Measured FSR: 0.194 nm

Estimated FSR from ring: 0.203 nm
Estimated FSR from III-V: 0.266 nm

➢ <u>Lasing with feedback from silicon photonics</u>





Gain section

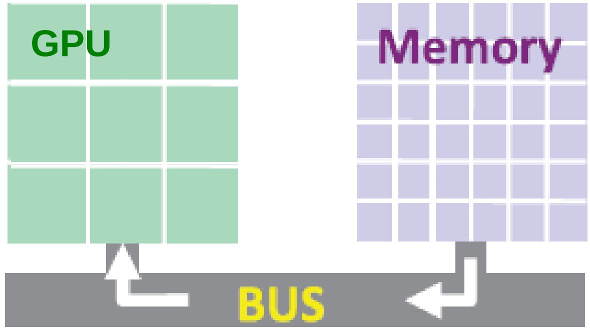
Outline

- **Chip-level 3D Technologies**
  - Introduction
  - 3DM for enhanced CMOS performance

- **Enhanced CMOS functionality**
  - CMOS Silicon Photonics
    - Monolithic laser integration
  - Novel hardware for Neural Networks
    - Resistive Crossbar Arrays
    - Photonic Synaptic Processor

- **Conclusions**

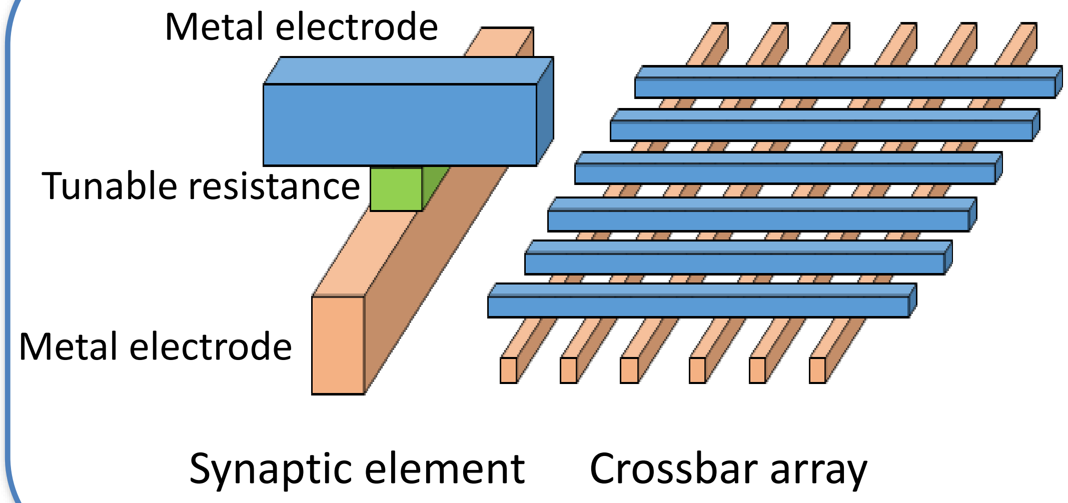
Neuromorphic hardware for big data analytics



- **Today’s status on deep neural networks**
  - Software based on Von Neumann systems
  - Training is the bottleneck – HPC required
  - GPU accelerators – processing memory bottleneck

- **Fast and efficient neural network data processing**

1. Analog approximate signal processing
2. Tight integration of processing and memory
   – 10’000x improvement using crossbar arrays



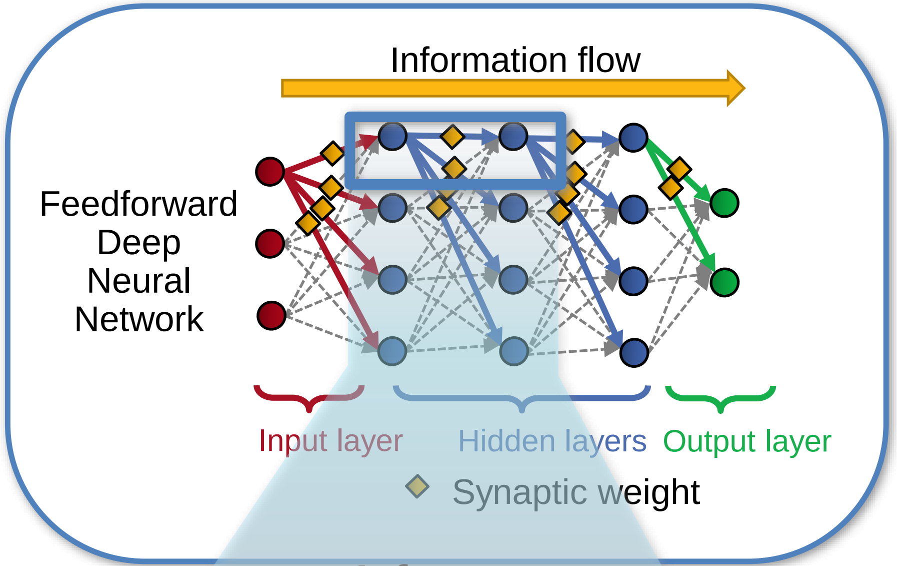
- **Hardware implementations**
  – *Electrical* crossbar arrays
  – *Photonic* crossbar arrays

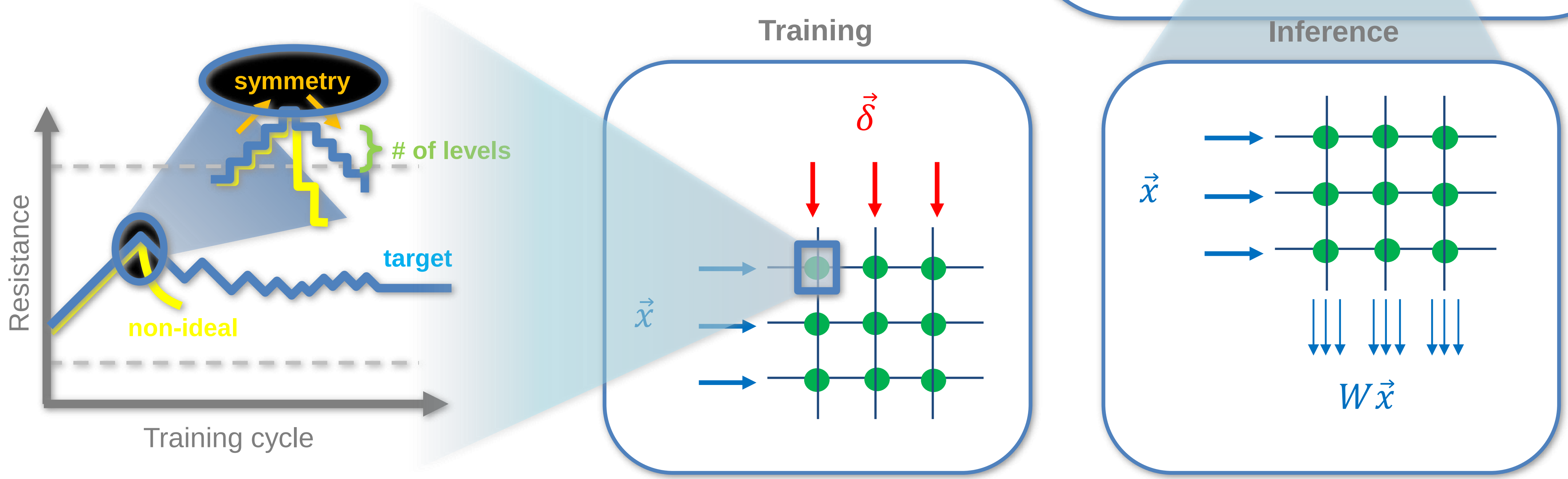
Accelerated learning: Analog crossbar arrays

Update weight proportional to signals on crossbar row and column

- Increase and decrease of weight
- Symmetric behavior for positive and negative updates
- High weight resolution (~1000 levels) required

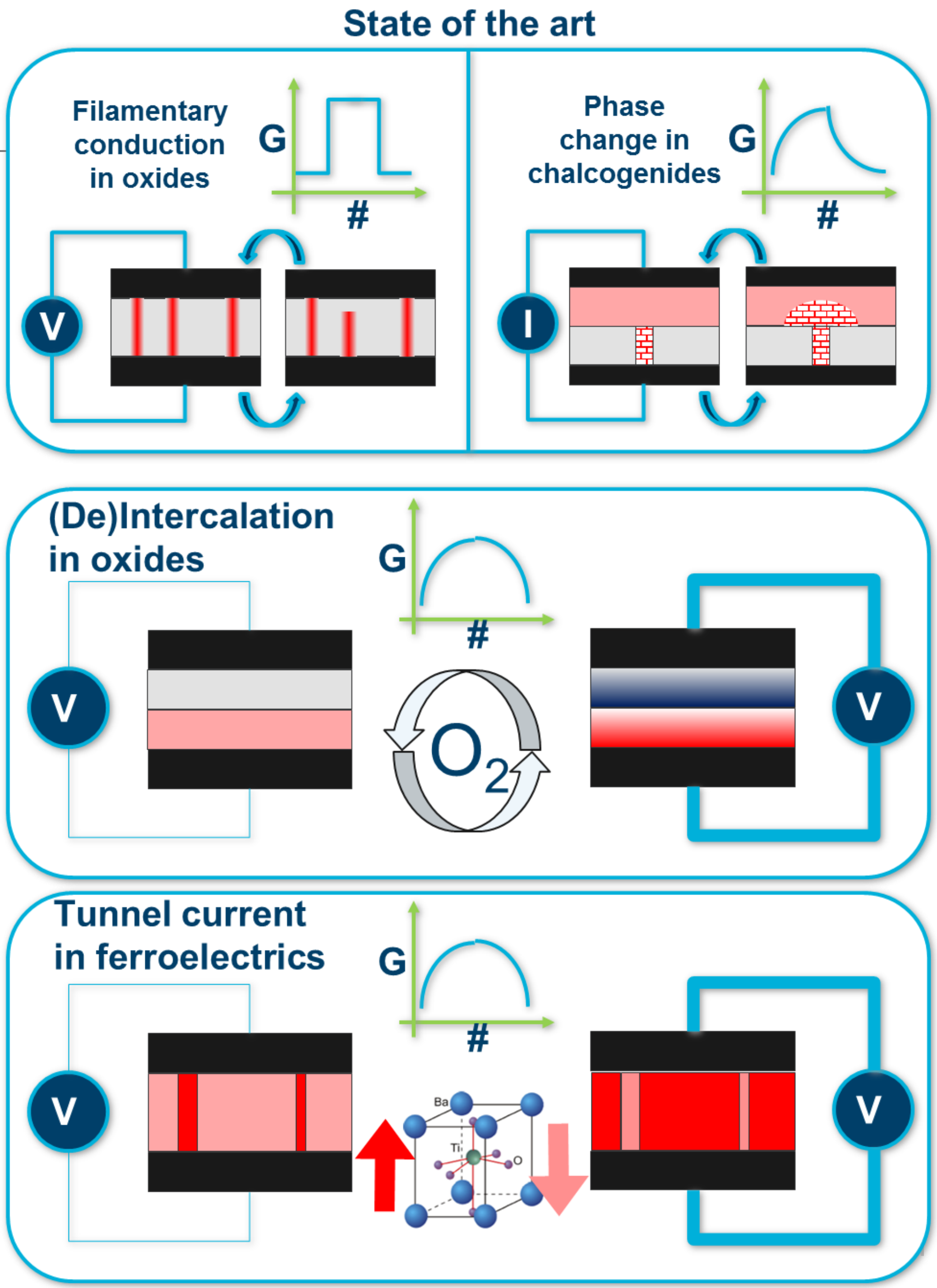
Physical challenge: Identify material systems fulfilling those requirements





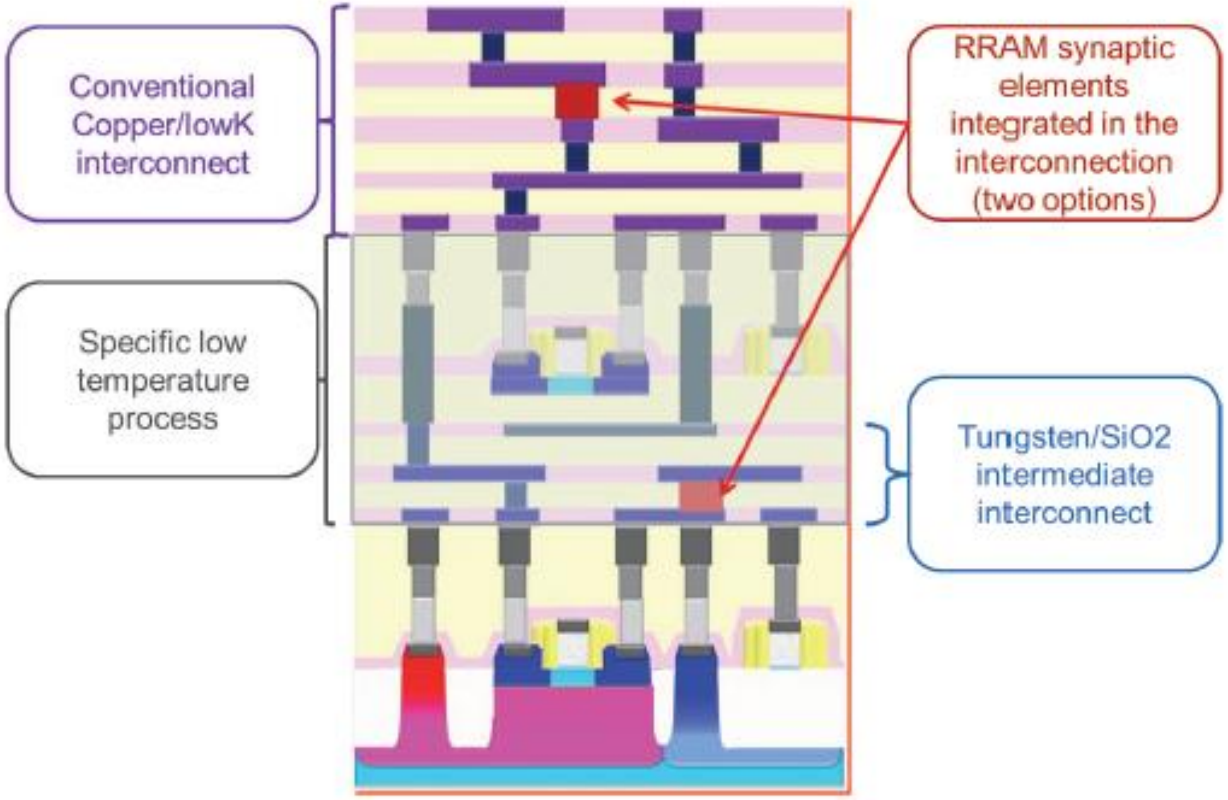
Materials for Resistive Crossbar Arrays

- **Synaptic functionality**
  - Multiply: Ohm's law; Accumulate: Kirchhoff's law
  - Weight update: *Tunable resistance*
- **Tunable resistance**
  - State of the art: filamentary oxides, PCM
    - Not analog enough or not symmetric enough
  - Search for *uniform* effects leading to ΔG with the appropriate resolution and symmetry
    - Electrochemical (de)intercalation
    - Control domain structure in ferroelectrics

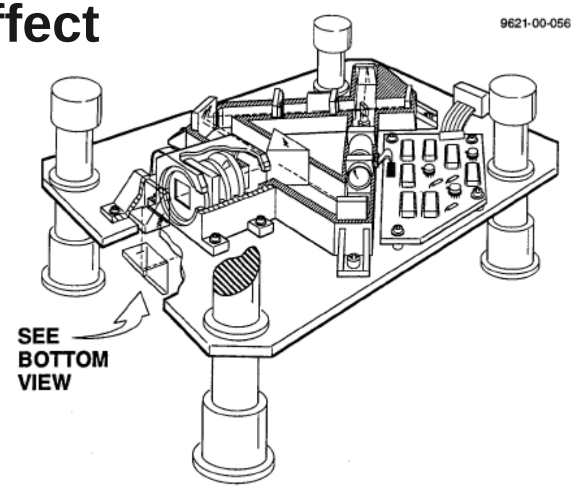


EU project NEURAM

- **Novel 3DM technology for neural networks**



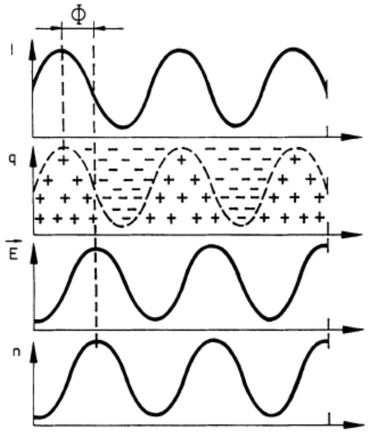
Photonic crossbar unit

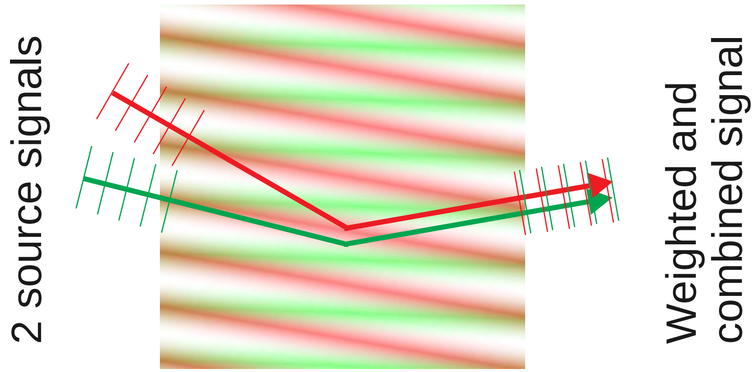


- **Alternative crossbar physical principle leveraging the photorefractive effect**
  - Demonstrated in 3D free space photonic neural networks in the 90s
    - i.e. Hughes Research Laboratories
  - New developments we can leverage
    - *Integrated* optic technology
    - Co-integration of new materials

- **Non-volatile weights applying the photorefractive effect**
    - Grating writing by interference of optical plane waves in an electro-optic material

      Strength proportional to product of the amplitudes of the writing beams
    - Written grating acts as the synaptic interface between plane optical waves



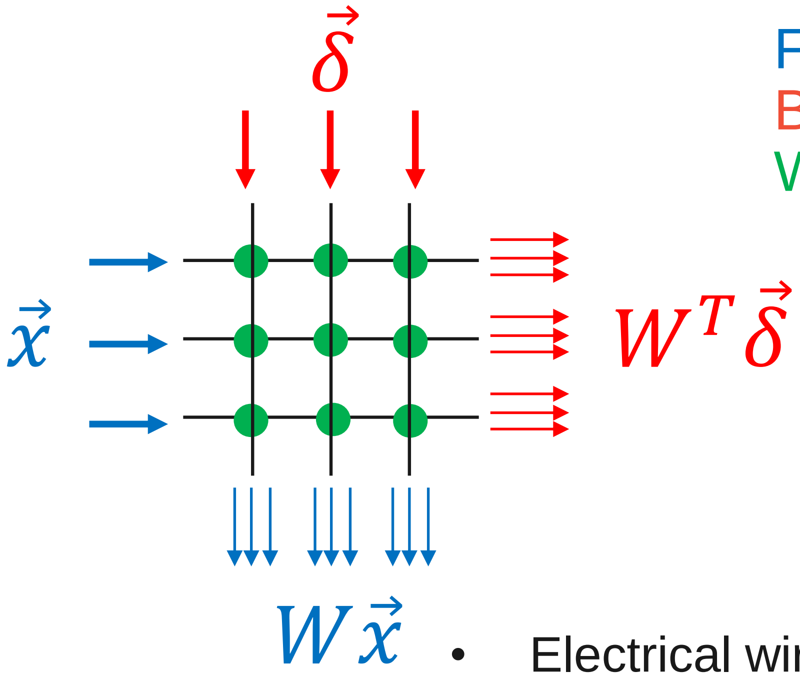


Photonic crossbar unit - operating principle

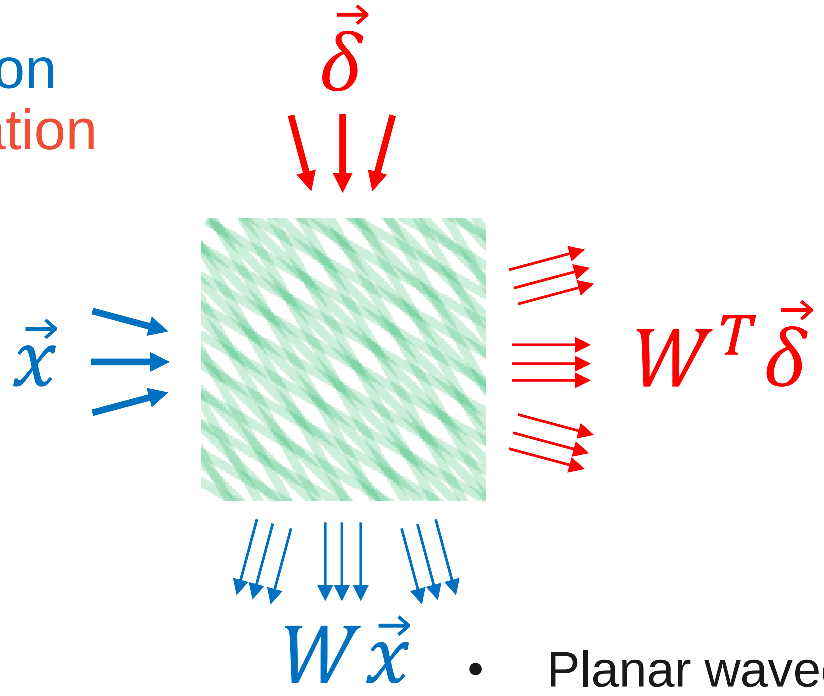
Electrical crossbar

Photonic crossbar

Forward propagation
Backward propagation
Weights



- Electrical wires
- Local weights
- Resistance tuning



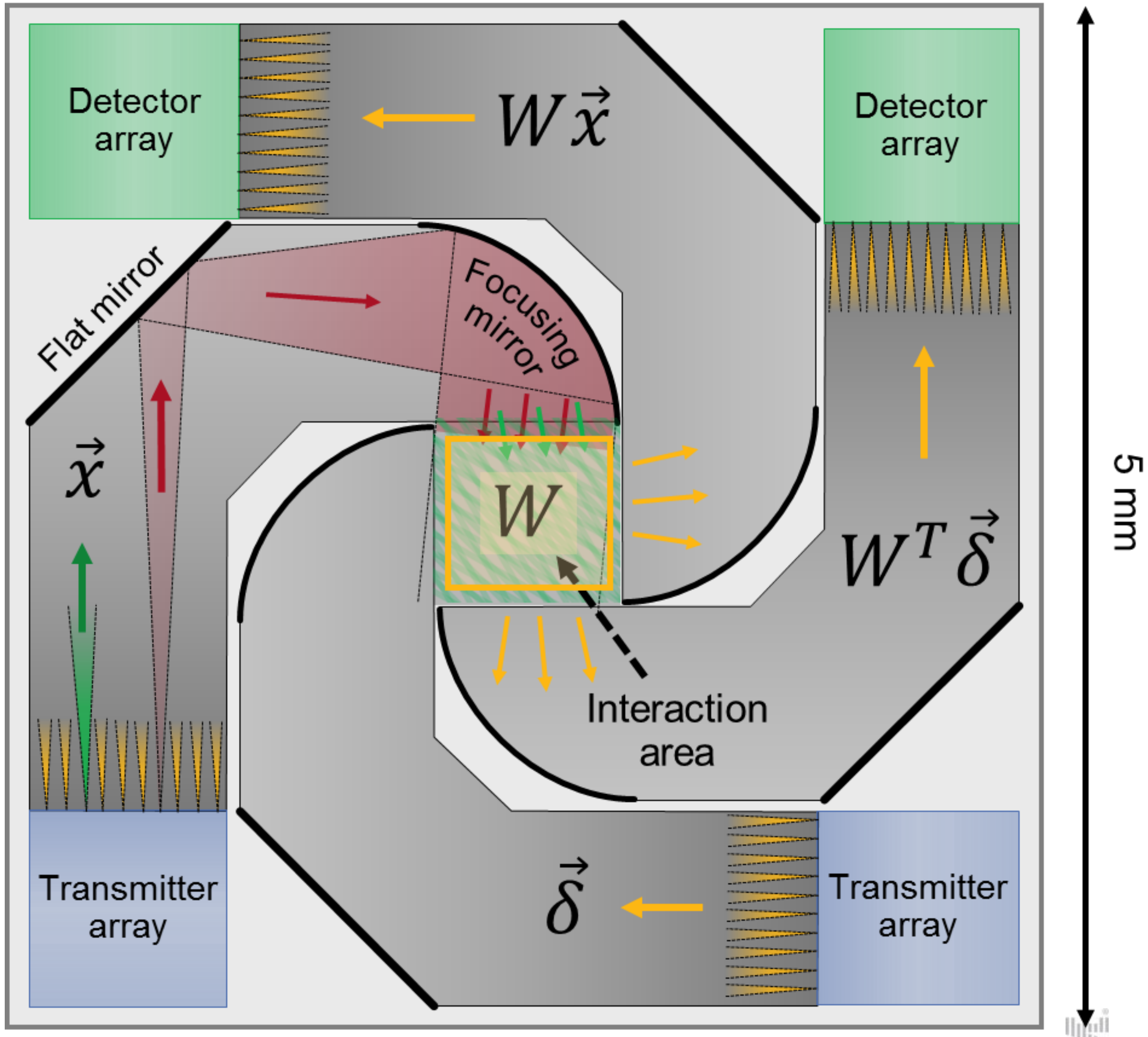
- Planar waveguiding
- Distributed weights
- Refractive index tuning

Writable photorefractive gratings provide the same functionality as the tunable resistive elements in a crossbar unit

Photonic crossbar unit - subsystem layout

Design study: 200 x 200 processing unit

- Weighting elements
  - Thin planar photorefractive slab on SOI, 1mm x 1mm
  - BTO or GaAs
- 200 Pre-synaptic electrical input signals
  - Optical power control through silicon photonic eo modulators
  - Plane optical waves under different angles in Si slab layer
- 200 Post-synaptic electrical output signals
  - Diffracted plane wave outputs are focused by Si mirror
  - Converted to electrical signals by a Si photonic detector array
- Weight adjustment
  - Photorefractive index modulation through interference pattern originating from the two transmitter arrays
- Scalability
  - A 200 x 200 processing unit fits in 5mm x 5mm
  - Theoretical capacity of interaction area is 2000 x 2000
  - Linear width and height scaling with # channels



Conclusions

- **CMOS technology performance enhancement**
- **3D Monolithic integration using III-V Materials**
  - Technology established

- **Enhanced silicon photonics**
  - Monolithically integrated lasers

- **Novel hardware for neural networks**
  - Co-integrated analog processing and memory
  - Electrical and Optical implementations of the synaptic processing unit

Acknowledgements

- **Collaborators in IBM**
  - Marc Seifried, Herwig Hahn, Gustavo Villares, Lukas Czornomaz, Folkert Horst, Daniele Caimi, Charles Caer, Yannick Baumgartner Daniel Jubin, Norbert Meier, Roger Dangel, Antonio La Porta, Jonas Weiss, Jean Fompeyrine, Veeresh Despandhe, Daniele Caimi, Lukas Czornomaz, Ute Drechsler
  - And many others

- **Co-funded by the European Union Horizon 2020 Programme and the Swiss National Secretariat for Education, Research and Innovation (SERI)**