Innovative Computing

Dr. Matthias Fertig

Professor of Computer Engineering

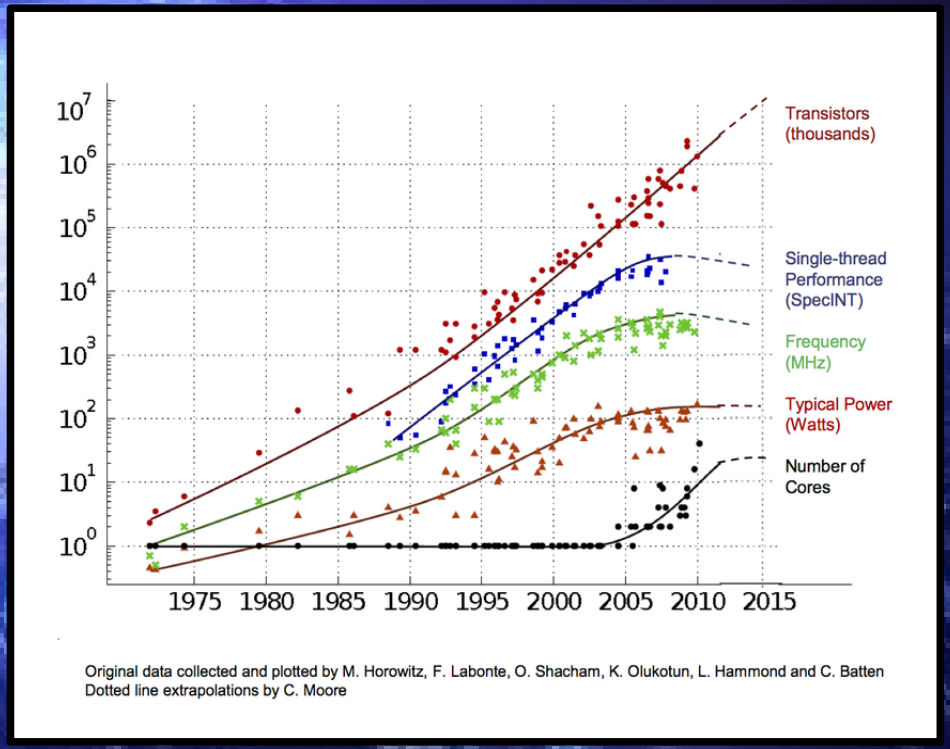
formerly CPM New Product Development
R&D Engineer at IBM Research & Development

Outline

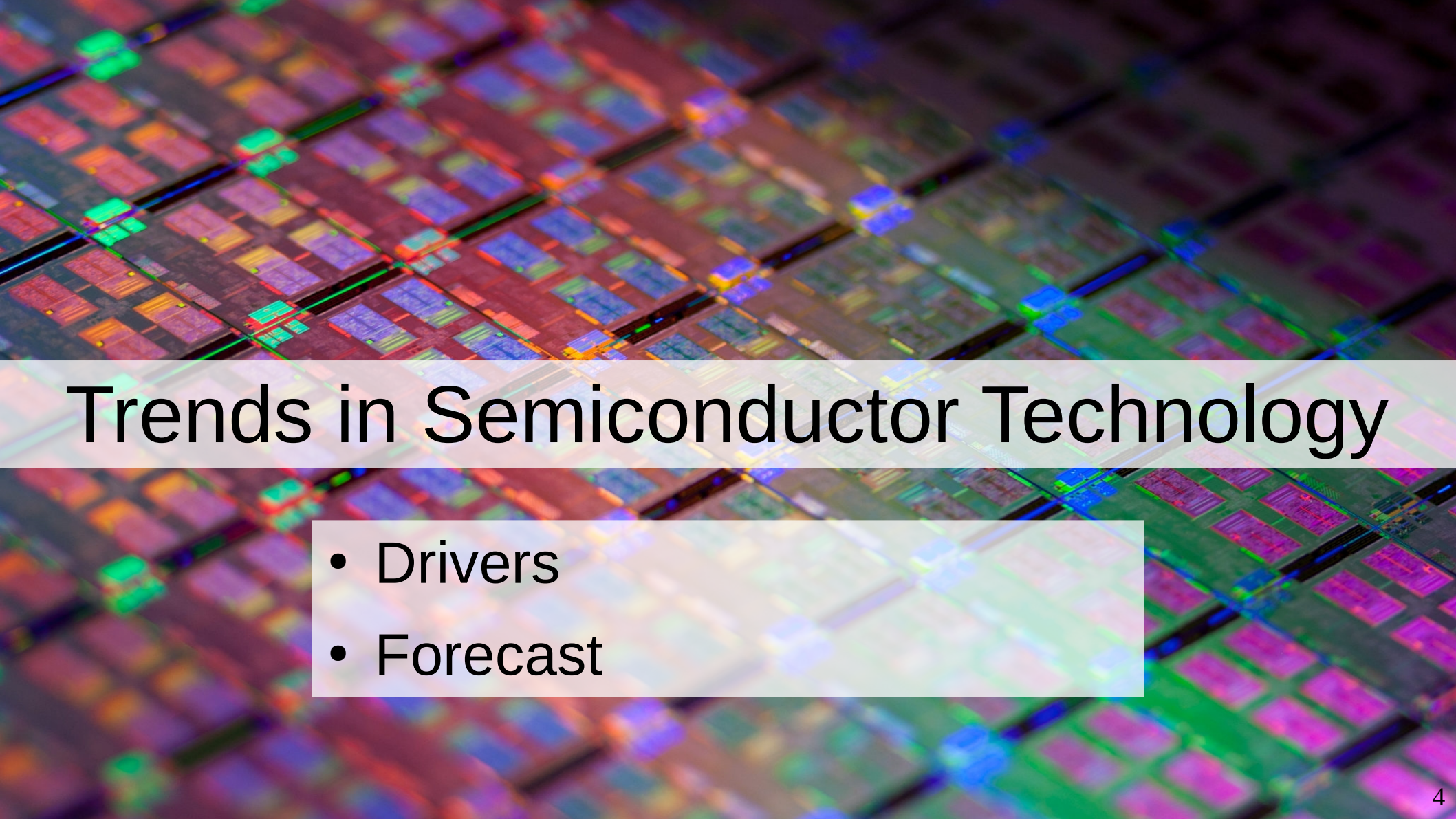
- Trends in Semiconductor Designs
- Semiconductor Technology
  - Drivers
  - Forecast
- Areas of Application & Innovation
  - Artificial Intelligence
  - Big Data & Internet of Things
- Research & Development Activities
  - Artificial Intelligence
  - Computing Technologies
  - Optical Interconnects
  - Advanced Materials
  - Near-/In-Memory Computing with Universal Memory Automata
- Vision & Next Steps

Trends in Semiconductor Designs

- Number of transistors doubles every 18 months (MOORE‘s law – G. Moore 1959)
- Power scales with transistor size (Dennard scaling)
  - Breaks in 2006
    - leakage current ► lower limit for power consumption
    - threshold voltage ► lower limit for supply voltage
- MOORE‘s law + Dennard scaling = KOOMEY‘s law
  - Broken since 2006 due to Dennard scaling
  - Performance per Watt does not scale with technology
- Saturation observed since 2006 in
  - Single-thread performance
  - Frequency
  - Number of cores per processor die
- Power and thermal are the limiting factors
  - Reduction of power density is needed



$$Performance \propto \frac{Energy}{Cycle} = \frac{Energy}{Instruction} \cdot \frac{Instruction}{Cycle}$$

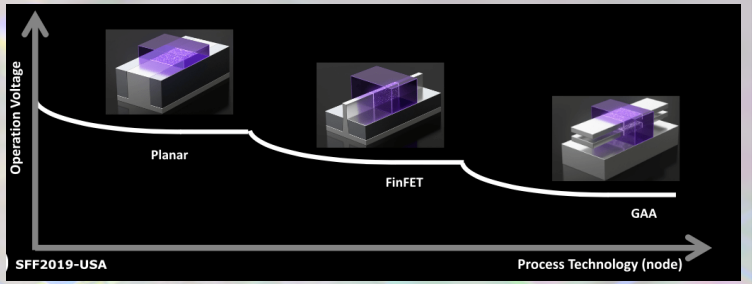


Trends in Semiconductor Technology

- Drivers
- Forecast

Technology – Drivers[1]

- Technology scaling performance boosters:
  - Advanced materials
    - High-K metal gate low leakage dielectrics
    - Strain induced charge mobility enhancement
    - ➔ Graphene Nanotubes
  - Innovative Device Architectures
    - FinFET until 2021
    - ➔ Lateral / Vertical Gate-All-Around (GAA) transistors
    - ➔ Operation voltage is going to decrease
  - ➔ Focus shifts from performance gain to power reduction!
- Performance-Power-Area (PPA) scaling requirements:
  - Performance: > 15% higher operation frequency
  - Power: > 35% less energy per switching at expected performance*
  - Area: > 35% less chip area
  - Cost: < 30% higher wafer cost



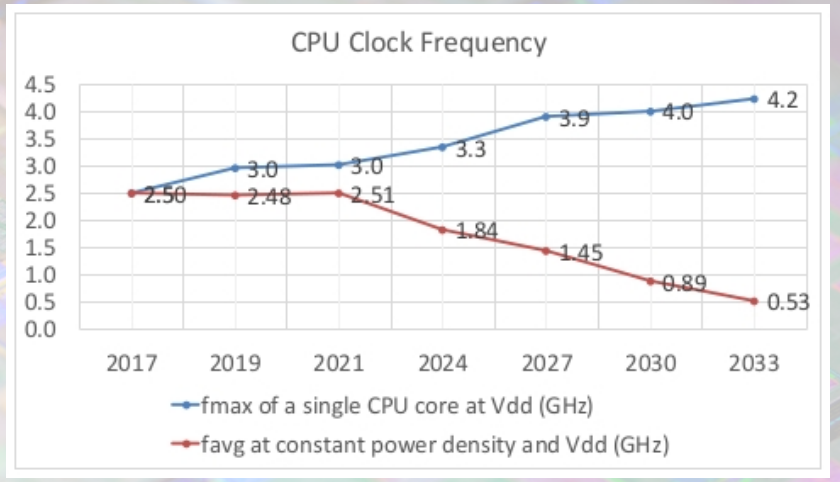
Operation voltage for Planar FET, FinFET and Gate-All-Around FET (Source: SFF2019-USA).

[1] IRDS - „The International Roadmap For Devices and Systems 2017“, IEEE 2018.
* 35% needed while 19% achieved in average from node to node according to [1].

Technology - Forecast[1]

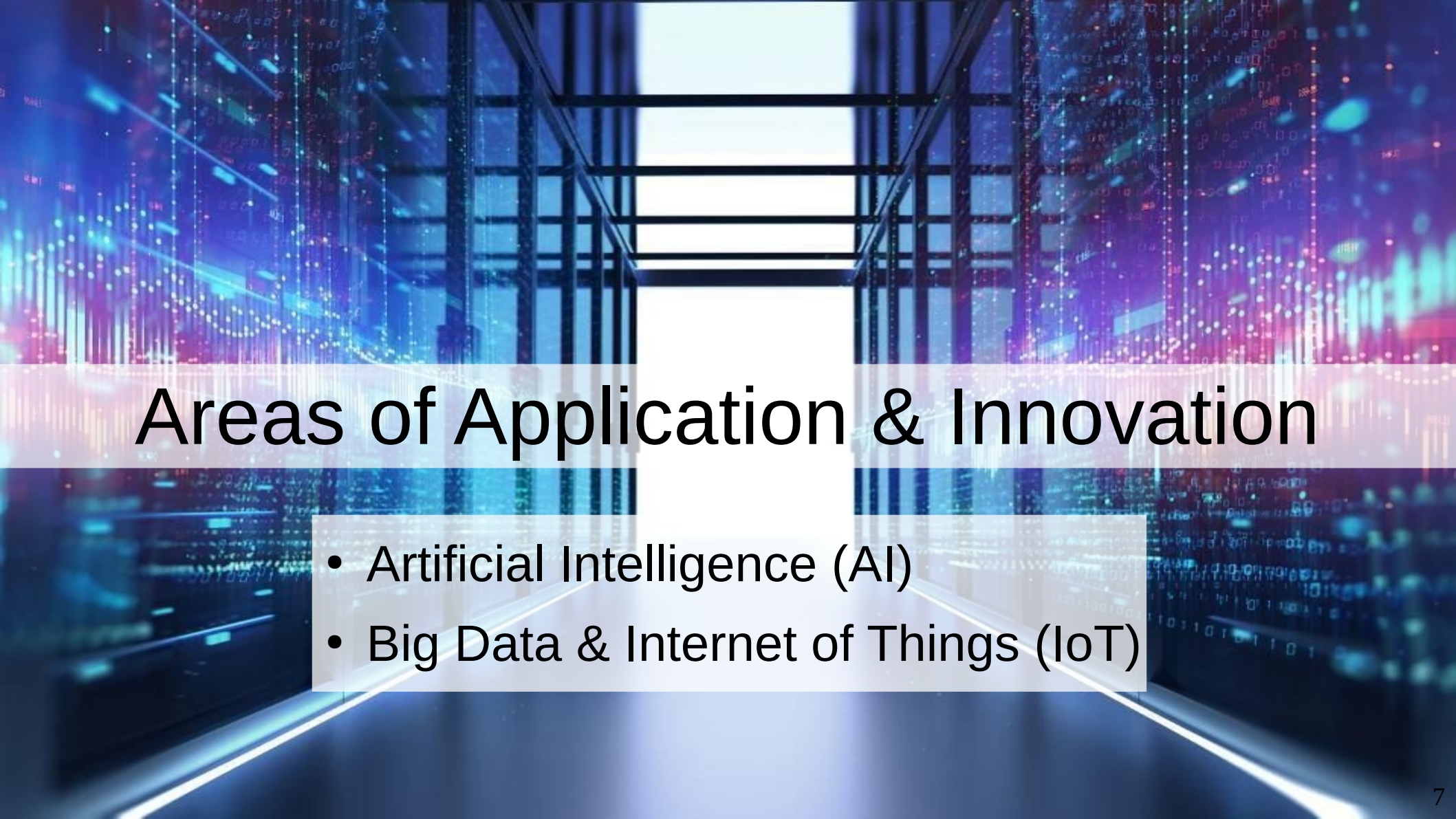
- Interconnect Technology Requirements[1,2]
  - Reduce electrical wire RC degradation
  - Increase electrical channel bandwidth & density
- System-level solutions[1,2]
  - Optical Chip-to-Chip Interconnects:
    - ► Photo diodes & Optical modulators
    - ► Waveguides & Wave Division Multiplexer
- Forecast (IRDS 2018)
  - Frequency saturation somewhere around 5GHz
  - Frequency degradation if power density stays constant and complexity continues to increase

CPU clock frequency scaling over time



CPU clock maximum frequency scaling at constant Vdd and average frequency scaling at constant power density and Vdd [1].

[1] IRDS - „The International Roadmap for Devices and Systems 2017“, IEEE 2018.
[2] IRDS - „International Roadmap for Devices and Systems 2016 Edition“ Outside System Connectivity White Paper



Areas of Application & Innovation

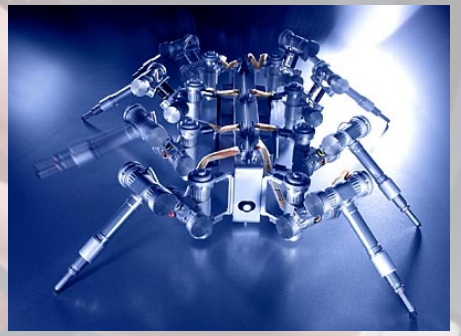
- Artificial Intelligence (AI)
- Big Data & Internet of Things (IoT)

Artificial Intelligence

- Deep Neural Networks & Convolutional Neural Networks
  - Millions of parameters form a huge design space ► adjust parameters to train network
  - Lots of matrix operations ► Fast Multiply-Add (dot-product) execution required
- HW-Accelerators & Algorithms ► Systolic arrays & tools for design space exploration

Autonomous (Mobile) Robotics

- Artificial Intelligence to increase level of autonomy
  - Using FPGAs or Hardware Accelerators
  - FPGA used in DARPA/NASA-funded project SCORPION[1]
    - Leg control & leg collision avoidance by path synchronization
    - Situation awareness by sensor fusion (ultra-sonic & tactile)



[1] F. Kircher, D. Spenneberg, M. Fertig et. al., **DARPA** grant number **N00014-99-1-0483** and **NASA** grant number **USRA 8008-003-002-01**, 2002

Big Data & Internet of Things

Big Data:

- Large sets of heterogeneous unstructured & distributed data
  - Cloud space
- Moving data to the processing nodes is expensive
  - Cloud computing

Internet of Things:

- Heterogeneous network with different type of nodes
  - Cloud server, smartwatch, car, fridge, ...
- High amount of communication between nodes
- Different areas of application, protocols & standards

- Requirements
  - Increased interconnect bandwidth & channel density
  - Increased Memory capacity, speed & reliability
  - Reduced power consumption
  - Security

→ Near-/In-Memory Computing[1] architectures
  - Data-centric approach of computing

[1] Singh, Boonstra et. al. „A Review of Near-Memory Computing Architectures: Opportunities and Challenges“, Eindhoven University of Technology, 2018

Research & Development Activities

Artificial Intelligence

German Ministry of Education and Research (BMBF):

03/2020 „Erforschung, Entwicklung und Nutzung von Methoden der KI in KMU“
01/2020 „Roboter für Assistenzfunktionen: Interaktion in der Praxis“
10/2019 „Künstliche Intelligenz in der zivilen Sicherheitsforschung“
04/2019 „Richtlinie zur Förderung von KI-Laboren in Deutschland“
04/2019 „Intelligentes Zusammenwirken von Mensch und Künstlicher Intelligenz“
04/2019 „Erklärbarkeit und Transparenz des Maschinellen Lernens und der KI“

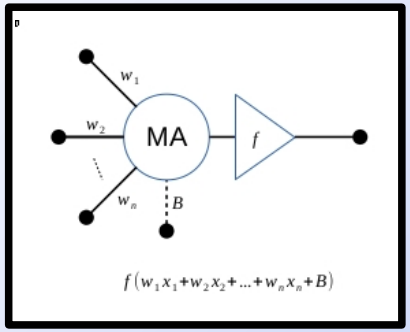
► 29 projects on „Maschinelles Lernen“ and „Algorithmenentwicklung“ funded

Horizon 2020 EU - „Work-Programme 2018-2020“:

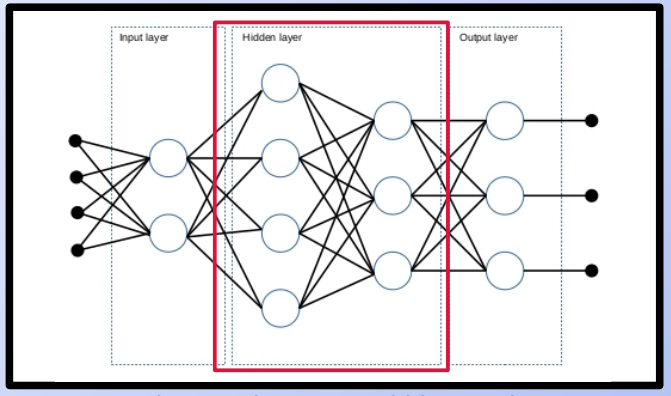
ICT-48-2020: Towards a vibrant European network of AI excellence centres
ICT-49-2020: Artificial Intelligence on demand platform
ICT-38-2020: Artificial intelligence for manufacturing
ICT-09-2019-2020: Robotics in Application Areas
ICT-10-2019-2020: Robotics Core Technology
ICT-26-2018-2020: Artificial Intelligence

* ICT - Information and Commmunication Technologies

Deep Neural Networks (DNN)



Perceptron - Multiply-Add (MA)

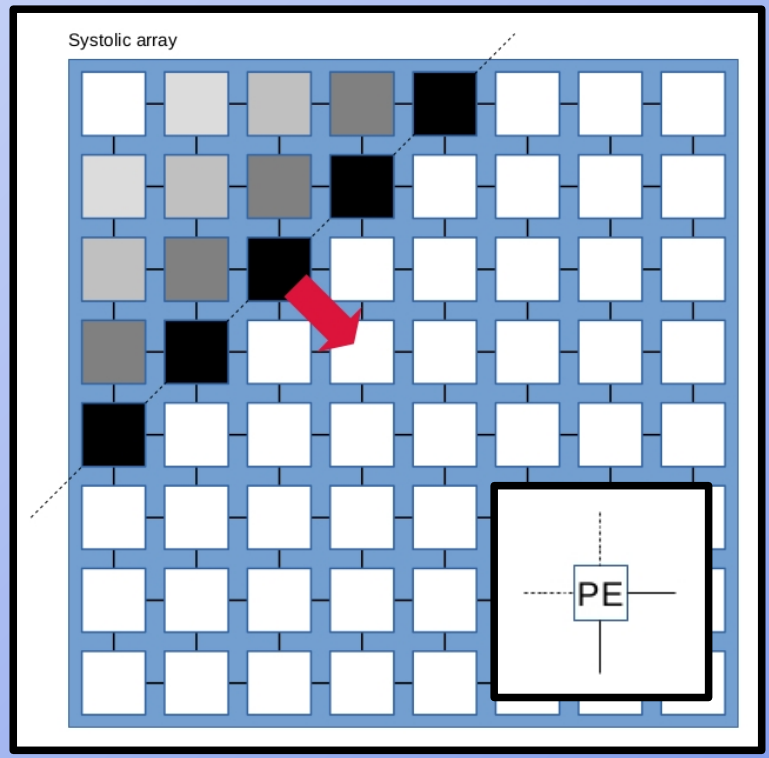


Deep Neural Network – Input- Hidden- and Output Layer

- DNN-characteristics
  - Series of recursive functions ► Design Space Explosion
  - Energy & Latency is critical ► High Performance per Watt

➔ Hardware-Accelerators ► Systolic Arrays[1]
  - 2D array of processing elements (PE)
  - Multiply-Accumulate (MAC) Instruction is key operation[2]

➔ Methodologies and tools for design space exploration

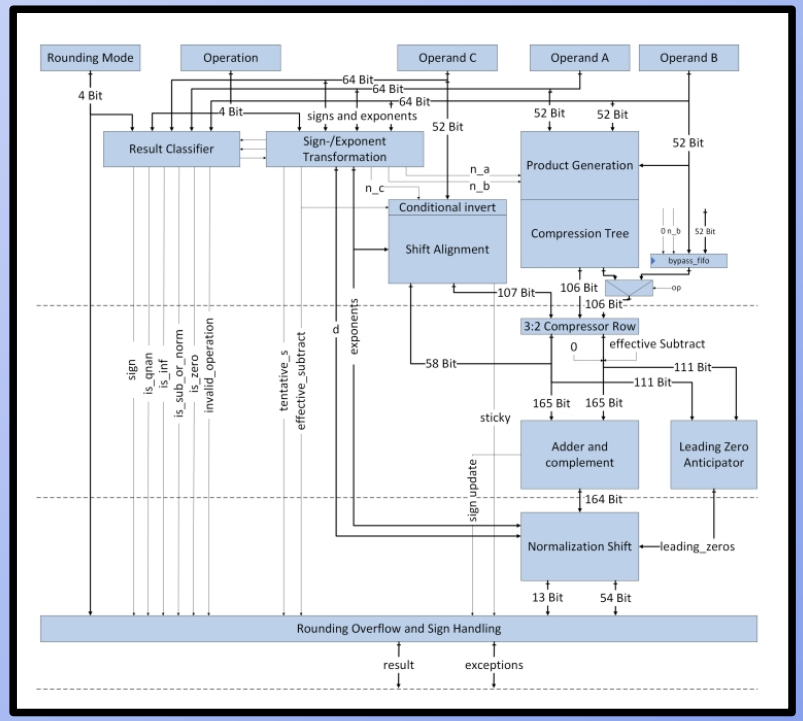


2D systolic array[1] with Multiply-Accumulate Processing Elements for (Deep) Neural Network Applications.

[1] Leiserson and Kung 1978, Carnegie Mellon University (CMU), inventors of the systolic array used in a corporation with Intel in the „iWarp“ chip in 1988
[2] F. Kaiser, „Design and Verication of a RISC-V Conform Fused Multiply-Add Unit“, Master Thesis at the Dpt. of Computer Architecture, Heidelberg Univ., 2018

Fused Multiply-Add Unit

- **RISC-V Double Precision Floating Point Unit[1]**
  - 3-Cycle Float Multiply-Add pipeline
  - Verified with **IBM FPGen[2]** framework against **Intel FMA Intrinsics[3]** & **Berkeley Softfloat**
- IEEE-754 Floating Point format
  - Is a convention and not a standard
  - Different results across platforms
  - Complicated special numbers (NaN, Inf, Denorm)
- → POSIT[4] format
  - Higher dynamic range and decimals accuracy
  - Higher speed, less energy & cost
  - 3-6x faster dot product[5] ► Systolic Array & AI



IEEE-754 Float Multiply-Add dataflow (Source: F. Kaiser)

[1] F. Kaiser, „Design and Veritifaction of a RISC-V Conform, Double-Precision Fused Mutiply-Add Unit“, Master Thesis, University of Heidelberg, 2018
[2] M. Aharoni, et. al., „FPgen - a test generation framework for datapath floating-point verification“, 8th IEEE Intl. High-Level Design Validation and Test Workshop
[3] https://software.intel.com/sites/landingpage/IntrinsicsGuide/#FMA
[4] Gustavson & Yonemoto, „Beating floating point at its own game: Posit arithmetic“, Supercomputing Frontiers and Innovations, 4(2):71 86, 2017
[5] D. Biancolin, J. Koenig, “Hardware Accelerator for Exact Dot Product”, ASPIRE Laboratory, UC Berkeley

Computing Technologies

German Ministry of Education and Research (BMBF) - High-Tech Strategie 2025

01/2020 „Richtlinie zur Förderung der Mikroelektronik-Forschung von deutschen Verbundpartnern im Rahmen des europäischen EUREKA-Clusters PENTA“
https://www.bmbf.de/foerderungen/bekanntmachung-2818.html

04/2019 „Zukunftssichere Spezialprozessoren und Entwicklungsplattformen (ZuSE)“

02/2019 „Mikroelektronik für Industrie 4.0 (Elektronik I4.0)“

Horizon 2020 EU - „Work-Programme 2018-2020“:

01/2020 - „Pan-European partnership in micro- and nano-electronic technologies and applications“ - PENTA ►PENTA is a EUREKA cluster managed by AENEAS

ICT*-54-2020: Blockchain for the Next Generation Internet

ICT*-01-2019: Computing technologies & engineering methods for systems of systems

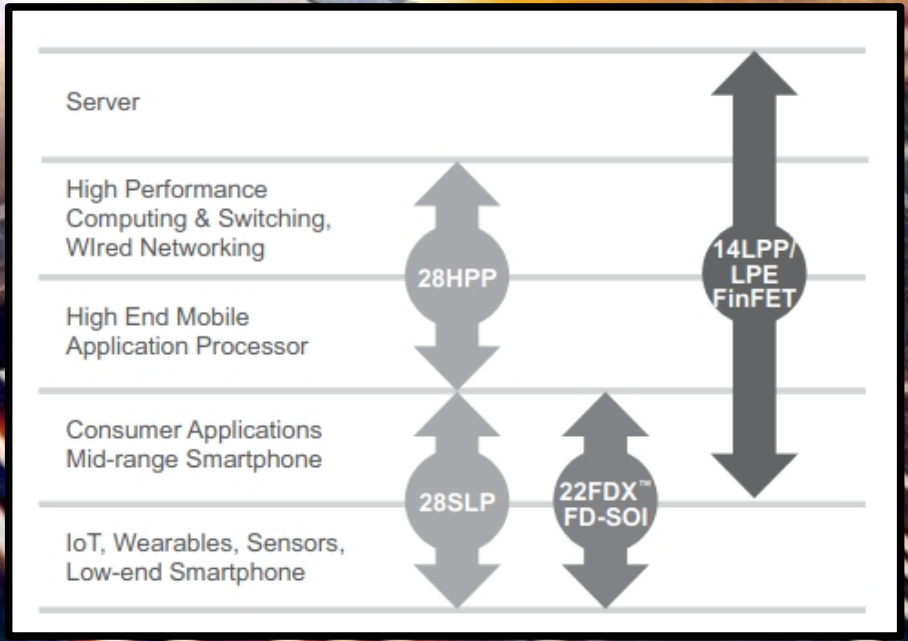
ICT*-06-2019: Unconventional Nanoelectronics

ICT*-07-2018: Electronic Smart Systems (ESS)

* ICT - Information and Commmunication Technologies

Semiconductor Technology

- High-performance to power-efficient System-On-Chip applications
  - Machine Learning & Artificial Intelligence
  - Internet of Things mobile processors
  - Networking
- Silicon Nano-Photonics support
  - GlobalFoundries 90GW technology
  - Based on 90nm technology
  - CMOS compatible

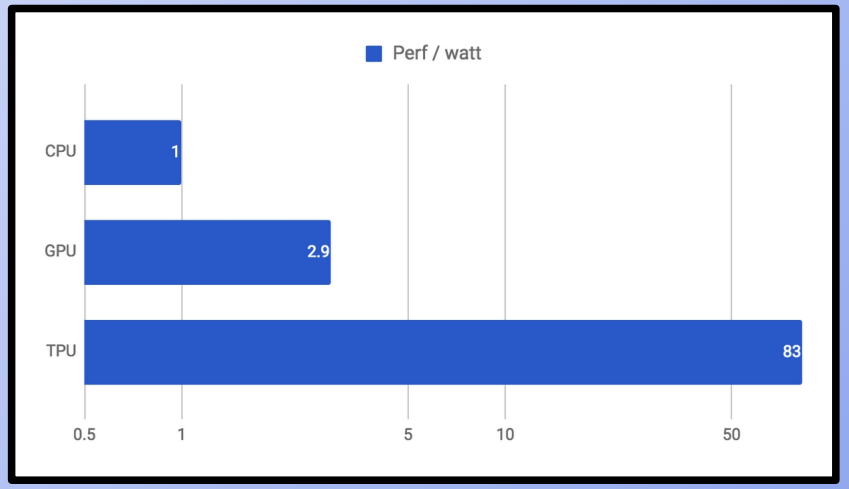


Source: Globalfoundries, Dresden.

Chip-manufacturers recognise the need for application-specific technologies!

RISC-V & Hardware Accelerators

- RISC-V – University of California Berkeley
  - Free Instruction Set Architecture ► no license fees
  - Optimized for efficient digital implementation
  - User extendable instruction set
- Application-Specific Hardware Accelerators
  - Machine Learning ►Google Tensor Flow (TPU)
- RISC-V with Accelerator support[2]
  - „Close“ vs. „Loose“ coupling



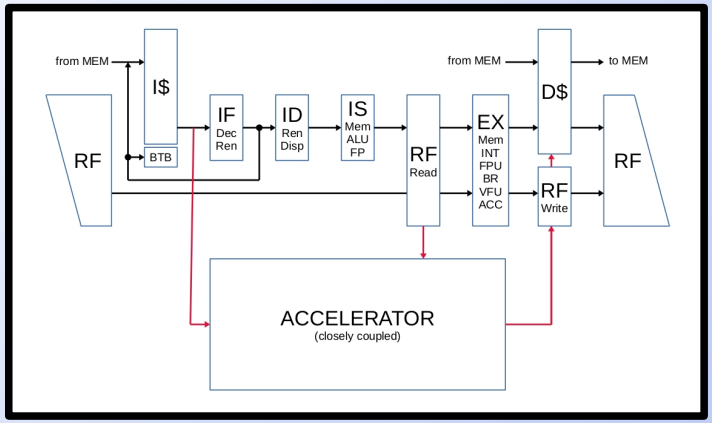
Source: Google

An open Instruction Set Architecture is a disruptive event!

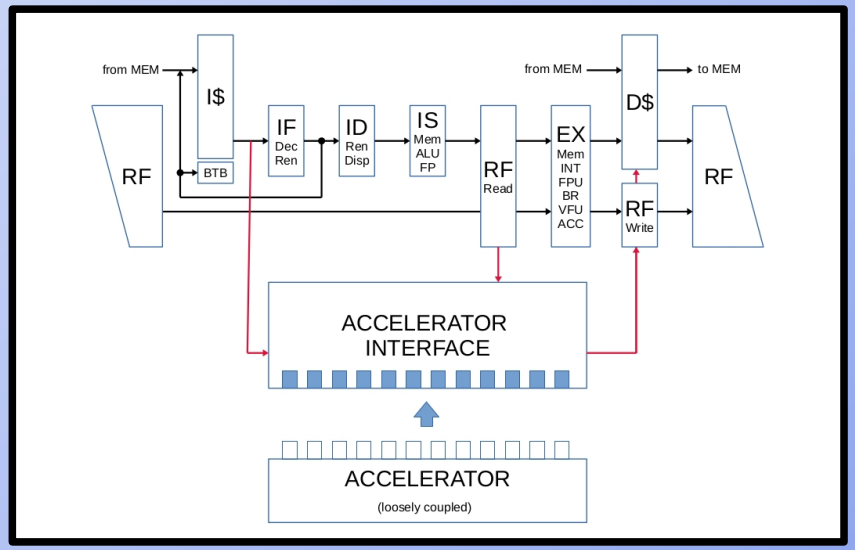
Hardware accelerators improve Performance per Watt ratio[1]!

[1] Timothy Prickett Morgan: Tearing apart Google's TPU 3.0 AI Coprocessor. In: The Next Platform. 10. Mai 2018
[2] U. Brüning, A. König, E. Suarez, M. Fertig et. al. „SEPIA – A Scalable Extendable Platform for Intelligent Applications“, BMBF project proposal 2019

RISC-V Microprocessor-Accelerator



Microprocessor pipeline with „closely“ coupled accelerator [1].



Microprocessor pipeline with „loosely“ coupled accelerator [1].

Standard Superscalar Microprocessor Pipeline

- Multiple Instruction Fetch, Issue & Dispatch
- Out-of-order execution
- Advanced Branch Execution
- Advanced Power Management

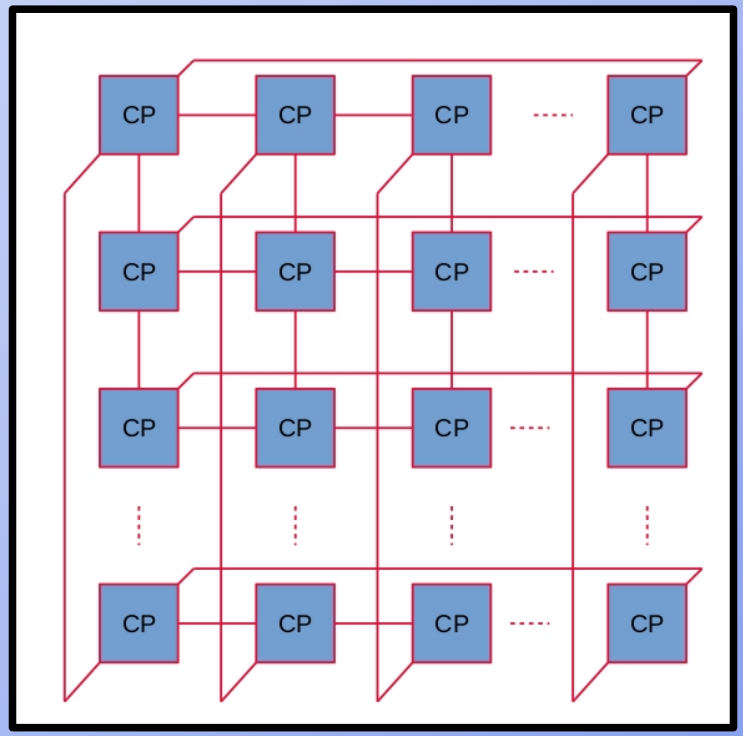
Application Specific Accelerator Support

- „Close“-Coupling vs. „Loose“-Coupling
- Generic Accelerator Interface

[1] U. Brüning, A. König, E. Suarez, M. Fertig et. al. „SEPIA – A scalable extendable platform for intelligent applications“, BMBF project proposal 2019

Optical Interconnected Transputer

- Transputer
  - First described in a 1978 paper by C. Hoare
  - Parallel Cluster ► 3D-Torus
    - ► Optimize number of node-to-node hops
- Electrical interconnects
  - ► Signal wiring is critical
  - ► Power, bandwidth & signal integrity issues
- Optical Interconnect Accelerator
  - ► Low power consumption
  - ► High data bandwidth
  - ► High channel integration density
- RISC-V Core with Optical Interconnect Accelerator



Transputer with Optical Interconnect Network.

[1] Principle of Operation developed 1978 by Inmos Corp, today ST Microelectronics
[2] U. Brüning, A. König, E. Suarez, M. Fertig et. al. „SEPIA – A scalable extendable platform for intelligent applications“, BMBF project proposal 2019

Silicon (Nano-)Photonics

German Ministry of Education and Research (BMBF):

06/2019 „Förderung von Vorhaben der strategischen Projektförderung unter der Beteiligung von Wirtschaft und Wissenschaft in den Bereichen Optik und Photonik“

01/2019 500 M€/5J „Photonik Forschung Deutschland“,

- https://www.photonikforschung.de/foerderung/foerderprogramm.html
- https://www.photonikforschung.de/projekte/explorative-forschung/foerdermassnahme/wivopro.html

Horizon 2020 EU - „Work-Programme 2018-2020“:

ICT-36-2020: Disruptive photonics technologies

ICT-37-2020: Advancing photonics technologies and application driven photonics components and the innovation ecosystem

ICT-03-2018-2019: Photonics Manufacturing Pilot Lines for Photonic Components & Devices

ICT-06-2019: Unconventional Nanoelectronics

ICT-42-2020: 5G PPP – 5G core technologies innovation

Optical Interconnect Advantages

- No Transmission Line Effects
- No Crosstalk and inductance issues
- Bandwidth is independent of distance
- Power is independent of (short) distance
- Architectural advantages:
  - No aspect-ratio* constrained performance
  - Reduced power dissipation
  - Increased channel density & bandwidth

$$RC_{el} = \rho_l \cdot c_l \cdot l^2$$

$$RC_{el,scaled} = \left(\frac{\rho_l}{s^2}\right) \cdot \left(c_l \cdot s^2\right) \cdot l^2$$

$$RC_{el,scaled} = \rho_l \cdot c_l \cdot l^2$$

$$RC_{el,scaled} = RC_{el}$$

$$B_{el} \propto \frac{A}{l^2}$$

$B_{el} = bandwidth\,[bits/s]$

$A = el.\ wire\ area$

$l = el.\ wire\ length$

Elecrical wire delays do not scale.

[1] „Silicon Photonics Trajectory Accelerates“, Intel Corp., Sept 22nd 2019
[2] „Silicon Photonics: High Speed Optical Connectivity for 5G“, Intel Corp. Sept 23rd 2019
[3] „Silicon Photonics: Data Center Connectivity“, Robert Blum, Intel Corp., Q3/2017
[4] „Silicon Photonics Passes the Tipping Point“, John Williamson, Intel Corp., Q3/2017
[5] Y. Vlasov, IBM Thomas J. Watson Research Center, Yorktown, 2008

* W/L tansistor aspect ratio affects integration density and so the minimum design pitch of electrical wires

Photonics Improvement Factors

Performance indicators & improvement factors[1]:

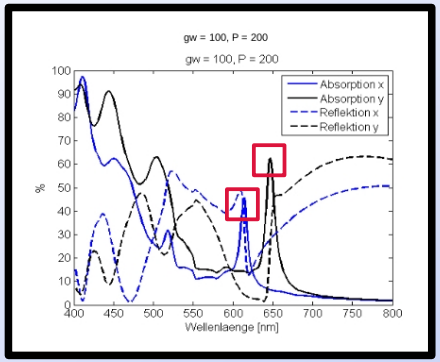
- Power ► 30x down
- Density ► 30x up
- Bandwidth ► 20x up
- Loss ► 10x down

Improvements equivalent to ~10 generations of CMOS![1]

CMOS is the lowest-cost platform![2]

[1] Y. Vlasov, IBM Thomas J Watson Research Center, Yorktown, 2008
Seek https://researcher.watson.ibm.com/researcher/view_group.php?id=5517 for more information
[2] C. Gunn (Vice President Technology at Luxtera), „CMOS Photonics Technology - Enabling Optical Interconnects“, 2017

Optical Grating Detector - Design



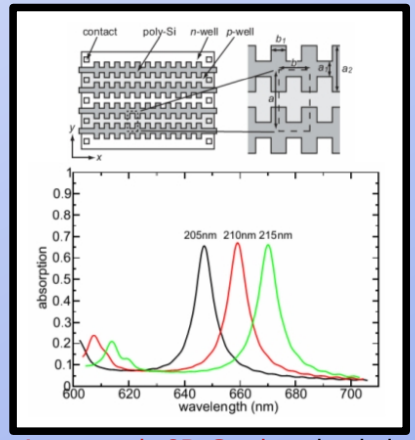
Symmetric 2D-Grating simulation.

$$k_{z,wg} = \frac{2\pi}{\lambda} \cdot n_c \cdot \sin\theta_t$$

$$k_{z,p}(m) = k_z - \frac{2\pi}{P} \cdot m, \quad m \in \mathrm{N}_0^+$$

$$P = \frac{\lambda}{n_c \cdot \sin\theta_t - n_{cl} \cdot \sin\theta_i}$$

Grating-Coupler theory (λ-θ correlation).



Asymmetric 2D-Grating simulation (~15nm channel separation) [1].

Joint Research Project[1] – IBM Böblingen, IBM ZRL & Heidelberg Univ.

- 32nm CMOS SOI ► only 80nm active detector silicon layer thickness
- 2D Grating Coupler optimized for TE- and TM-modes
- Coupling of modes m at angle-of-incidence θ into „parasitic“ waveguide
- λ to θ correlation ► sensitivity to angle of incidence
  ► high accuracy in semiconductor technology required
- Phase match condition to be met ► $k_{z,wg} = k_{z,p}$

Electromagnetic Simulation needed for optimization of grating geometry

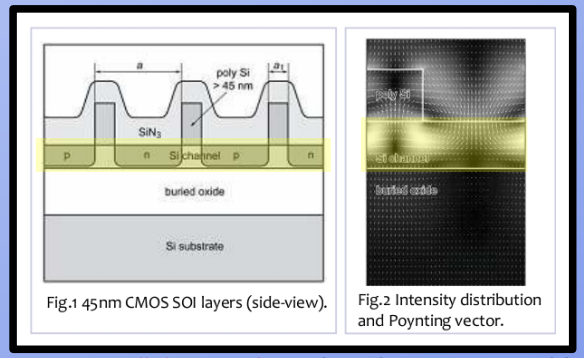


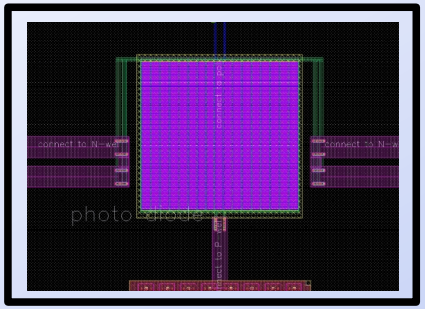
Fig.1 45nm CMOS SOI layers (side-view).

Fig.2 Intensity distribution and Poynting vector.

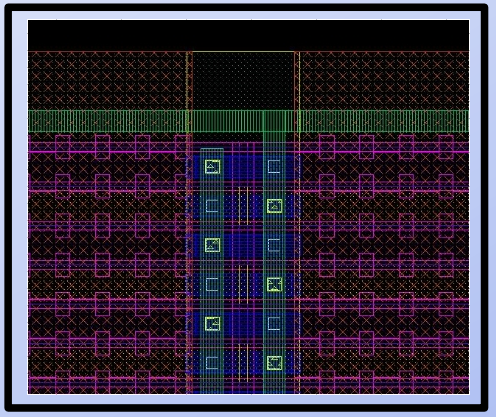
Resonant light-trapping w/ grating on waveguide. Simulation by B. Trauter & K. Brenner, 2007.

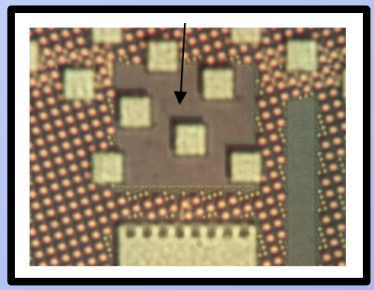
[1] M. Fertig et. al. „Polarization-Independent Photo-Detectors with Enhanced Responsivity in a Standard SOI CMOS Process“, IEEE JLT, 2009

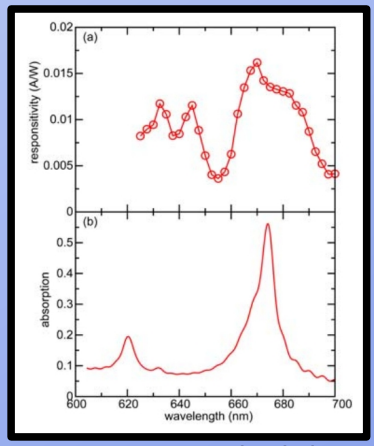
Optical Grating Detector - Results



Detector layout (Cadence Virtuoso)


2D-grating (Cadence Virtuoso)


Metal fill blockage (IBM)



Measurement vs. simulation of grating detector [1].

- Integrated Optical Detector[1]
  - Grating-Coupler[1,2,3] ► 11x absorption rate (6% to 66%) ► 8 patents pending
  - Good model to hardware correlation ► coincident absorption peaks
  - Responsitivity 0.56 A/W expected ► metal fill blockage layers
- Experiment confirmed by [3] in 2018

➔ Advanced Materials to achieve better responsitivity

► Germanium, Silicon Germanium, Gallium Arsenide ...

[1] M. Fertig et. al. „Polarization-Independent Photo-Detectors with Enhanced Responsivity in a Standard SOI CMOS Process, IEEE JLT, 2009
[2] M. Fertig et. al. **US Patents No US9219177B2, US8912616B2, US8543961B2, US8299556B2, US8225241B2, US7978983B2, US7880207B2, US7821091B2**
[3] S. Nambiar, P. Sethi, S. K. Selvaraja, „Grating-Assisted Fiber to Chip Coupling for SOI Photonic Circuits“, Journal of Applied Sciences, 13 July 2018

Advanced Materials

German Ministry of Education and Research (BMBF):

MetaTune - Elektronisch und optisch abstimmbare Metamaterialoptiken für Thz-Technologie
Nano-Film – Photonische Nano-Filme mit umfassender optischer Funktionalität
GraPhoTrans – Integrierter photonischer SiGe-Graphen 10Tb Transceiver
MILAS – Monolithisch integrierter Laser aus III-V-Nanodrähten für Silizium-Photonik

Horizon 2020 EU Program - CORDIS:

GRAB-A-META - „Graphene Based Active Metamaterials“
PHOME „Photonics Metamaterials: From Basic Research to Applications“
ICT-36-2020: Disruptive photonics technologies
ICT-02-2018: Flexible and Wearable Electronics
ICT-03-2018-2019: Photonics Manufacturing Pilot Lines for Photonic Components and Devices
ICT-07-2018: Electronic Smart Systems (ESS)

https://ec.europa.eu/research/industrial_technologies/pdf/metamaterials-brochure_en.pdf

Advanced Materials

- Advanced Materials ► Artificial Materials ► Metamaterials
  - „Artificial Materials“ exhibit special capabilities not observed for „Natural Materials“
  - Involves nano-technology and micro-fabrication
  - Metamaterials ► Negative Index Materials (NIMs)
- Cross-Fertilization of Increasing Computing Power and Advanced Materials
  - Computing Power can support the development of Advanced Materials
  - Advanced Materials can support the increase of Computing Power
- Economic potential of Advanced Materials
  - „Global Nanomaterials Market Set for Rapid Growth ► $16,8B by 2022“[1]

[1] Source: www.semiengineering.com

Negative Index Materials[1] 1/2

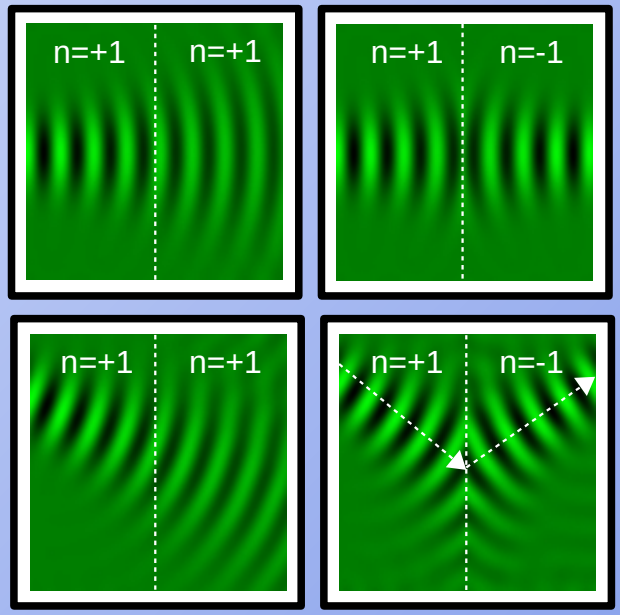
- Postulated by V. Veselago[1] (1968)
- Pioneering work by Pendry & Smith[2] (2004)
- Propagation constant is $k = n(\omega) \cdot \frac{\omega}{c} = k' + j \cdot k''$

$$k = k' + j\,k'' = \frac{\omega}{c}\sqrt{(\epsilon' + j\,\epsilon'') \cdot (\mu' + j\,\mu'')}$$

$$k = \frac{\omega}{c}\sqrt{\epsilon'\mu'} \cdot \left[1 + j\,\frac{1}{2} \cdot \left(\frac{\epsilon''}{\epsilon'} + \frac{\mu''}{\mu'}\right)\right] = k_0 \cdot n \cdot \left[1 + j\,\frac{1}{2} \cdot \left(\frac{\epsilon''}{\epsilon'} + \frac{\mu''}{\mu'}\right)\right]$$

- A changed sign of k' does not automatically change the sign of k'' ► amplification and refraction are independent
- $k' < 0$ ► **Material Lens**
- $k'' < 0$ ► **Quantum Amplification**

Simulation with VWPM[3], phase-inverted by negative index of refraction:



[1] V. G. Veselago: The electrodynamics of substances with simultaneously negative values of ε and μ. In: Usp. Fiz. Nauk. Band 92, 1964, S. 517–526.
[2] D. Smith, J. Pendry et. al., „Metamaterials and Negative Refractive Index“, Science 06 Aug 2004, Vol. 305, Issue 5685, pp. 788-792
[3] M. Fertig, „“The Vector Wave Propagation Method (VWPM)“, PhD thesis at the Department of Optoelectronics, University of Heidelberg, 2011 extended to negative index propagation by phase inversion. Fresnel coefficients not adapted to magnetic case yet.

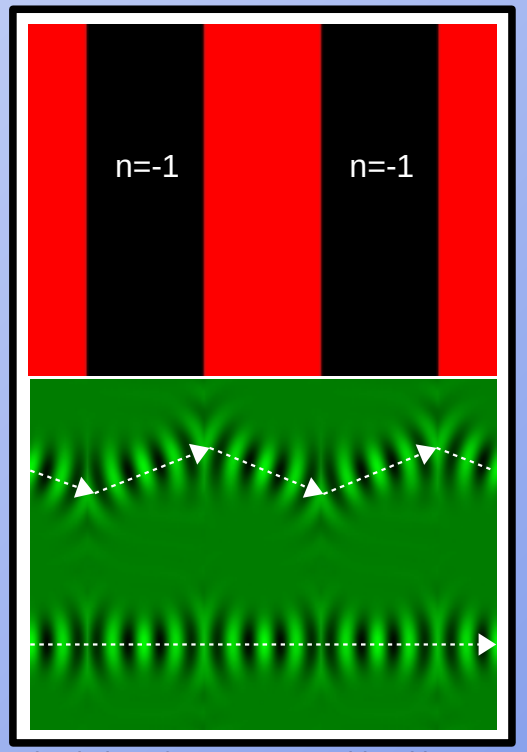
Negative Index Materials[1] 2/2

Two opportunities for Optical-Interconnects:

- Negative Index Waveguide
- Material Lens

Next steps:

- Develop NIMs at Si-Photonic λ ► needs nm design pitch[1]
  - ► **CMOS Technology**
- Computational methods to consider magnetic approach
  - ► **Computational Optics[1,2]**



Simulation of a NIM-waveguide without reflections with a phase-inverted VWPM[2,3].

[1] Fertig, „“The Vector Wave Propagation Method (VWPM)“, PhD thesis at the Department of Optoelectronics, University of Heidelberg, 2011
[2] Fertig, Brenner, „“The vector wave propagation mehdot (VWPM)“, Journal of the Optical Society America A, Vol 27, No. 4, pp. 709-717, 2010
[3] Veselago, Hafner, „Negative Refractive Index Materials, Journal of Computational and Theoretical Nanoscience, Vol. 3, 1-30, 2006

Big Data & Internet of Things

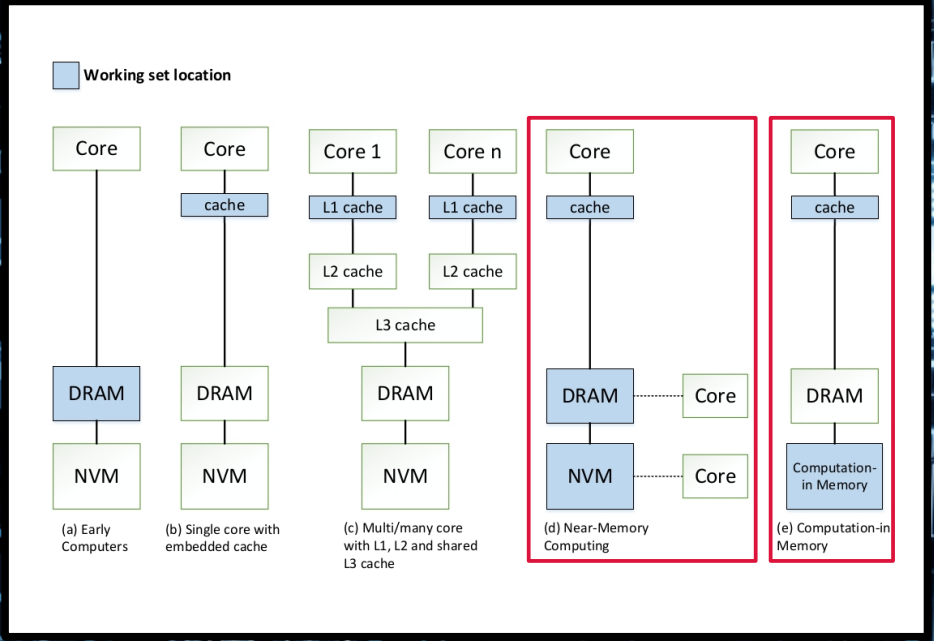
German Ministry of Research and Eduation (BMBF):

06/2019 „Neue Sicherheitstechnologien für das Internet der Dinge“
… funding for „Berlin Big Data Center“ & „Competence Center for Scalable Data Services“

Horizon 2020 EU Program CORDIS:

ICT-40-2020: Cloud Computing: towards a smart cloud computing continuum
ICT-11-2018-2019: HPC and Big Data enabled Large-scale Test-beds and Applications
ICT-51-2020: Big Data technologies and extreme-scale analytics
ICT-13-2018-2019: Supporting the emergence of data markets and the data economy
ICT-15-2019-2020: Cloud Computing
ICT-16-2018: Software Technologies
ICT-56-2020: Next Generation Internet of Things
ICT-24-2018-2019: Next Generation Internet - An Open Internet Initiative

Big Data

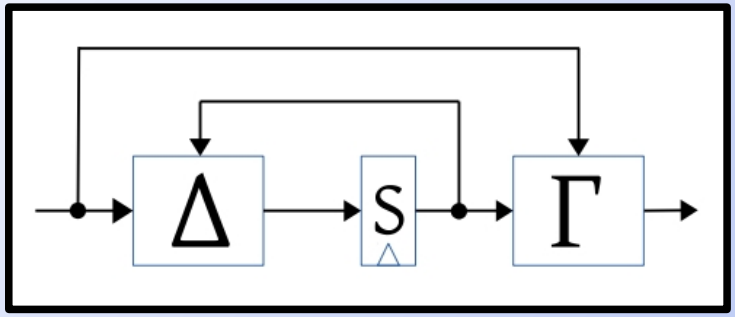


Conventional and Near-/In-Memory Compute Architectures [1].

- Moving large data sets is expensive
  - Move the instructions
- Instruction and data reverse roles
  - Data fetch instruction
- Data-centric compute model
- Near-/In-Memory Compute Architectures:
  - Near-Memory Computing
    - Cores attached to DRAM or NVM
  - In-Memory Computing
    - Computation in memory

[1] Singh, Boonstra et. al. „A Review of Near-Memory Computing Architectures: Opportunities and Challenges“, Eindhoven University of Technology, 2018

Universal **M**emory **A**utomata



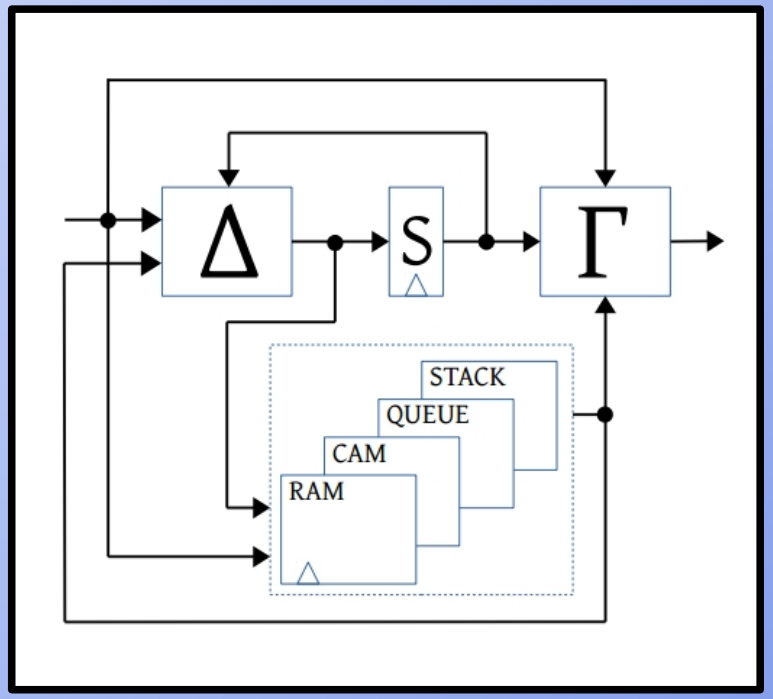
Finite State Machine Architecture (Mealy).

UMA[1] – A Data-Centric Compute Model

- RAM, Queue, CAM, Stack, ...
- Used to implement a **MESI Cache Coherency Protocol** (2018 Bachelor Thesis[2] – awarded the „MLP Leistungspreis“)
- Semi Custom Design Flow ► Near-/In Memory Architecture

► Advanced Memory Access Paradigms

RD/WR matrix row, column, diagonal elements



Universal Memory Automaton [1].

[1] M. Fertig „Universal Memory Automaton“, MPC Workshop Proceedings, Mannheim University of Applied Sciences, Mannheim, 2020
[2] P. Zyska, „Automated Verilog Code Generation of a Cache Coherency Snooping Protocol with Enhanced Push Down Automation“, Bachelor Thesis, Konstanz University of Applied Sciences, Konstanz, 2018

Vision ...

Transputer w/ Optical Interconnect System built on RISC-V Accelerators

Cross-functional Research & Development

- System Architectures (Free ISA, Application-Specific Accelerators, …)
- Software & Hardware Algorithms (Deep & Machine Learning, floating-point formats, …)
- Technology Support (Semiconductor Technology, Silicon Nano-Photonics, …)
- Material Science (Advanced Materials, …)

Project execution in **cross-functional global teams**

- Hard-skills in relevant fields of expertise!
- Soft-skills and inter-cultural competence!

… Next Steps

Development Activities

- ► Microprocessor development ► based on a free Instruction Set Architecture ► RISC-V
- ► Appl.-Specific Accelerator Interface development ► Fast Network Interface, Memory, Systolic Array
- ► Advanced Floating-Point formats ► POSITs ► RISC-V Floating Point Unit development
- ► Advanced design methodologies for Near-/In-Memory Computing ► Universal Memory Automaton

Research Activities

- ► Silicon Photonic Optical Interconnects ► Silicon Photonic Packages (Global Foundries, AMF*)
- ► Silicon Photonic devices & Metamaterials ► Computational Optics ► FDTD, VWPM & RCWA
- ► Advanced Materials ► Negative Index Materials ► Computational Optics & CMOS

Upcoming Funding Opportunities (BMBF):

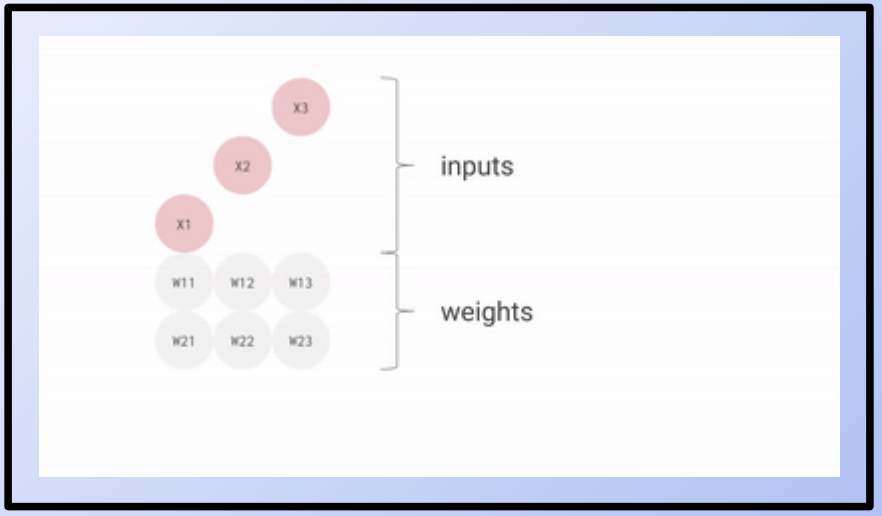
- ► Vertrauenswürdige Elektronik (ZEUS)
  www.bmbf.de/foerderungen/bekanntmachung-2888.html
- ► Erforschung, Entwicklung und Nutzung von Methoden der Künstlichen Intelligenz in KMU
  www.bmbf.de/foerderungen/bekanntmachung-2876.html

* Advanced Micro Foundry, Singapore, www.advmf.com – Integrated Photonics solutions for 400Gbps

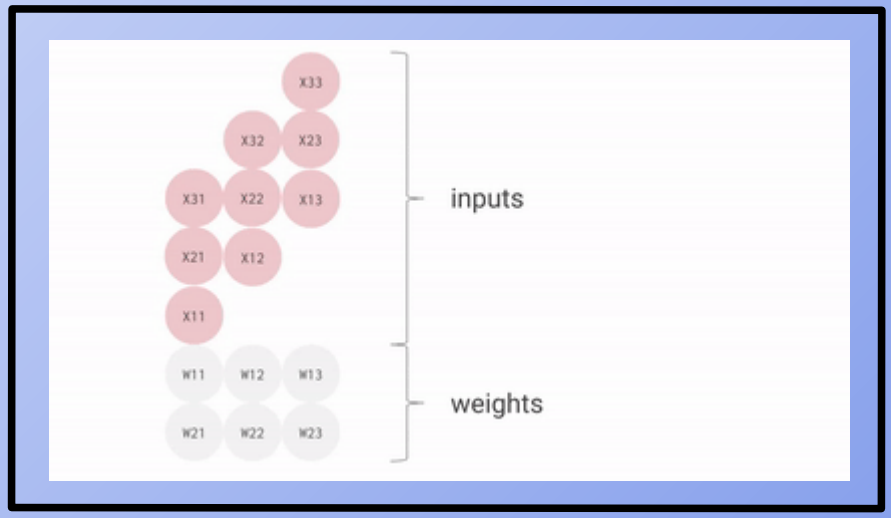

Thank you!

Backup Slides ...

Systolic Array
Principle of Operation[1]



Matrix-Vector-Multiplication with Systolic Array

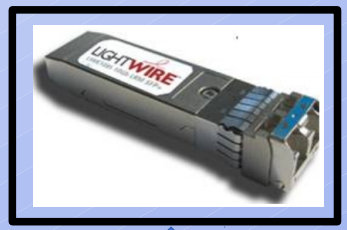


Matrix-Matrix-Multiplication with Systolic Array

[1] An in-depth look at Google's first Tensor Processing Unit (TPU)
https://cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

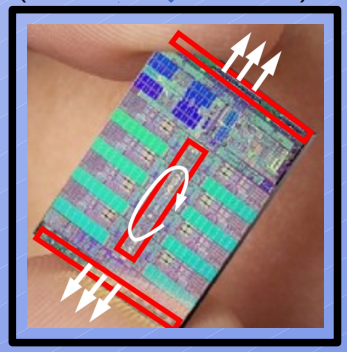
Optical Interconnect Talks

- **Intel Optical Interconnect Technology** – 2017 to 2020
  - Wafer-level integration of lasers achieved with bonding Indium Phosphide instead of post-processing ► cost reduction for packaging
  - Silicon Photonic Technologies now pushed by GOOGLE, Amazon, Facebook or Microsoft ► >400G at highly integrated form factors
  - „Silicon Photonics passes the tipping point“, J. Williamsom, Intel 2017 ►~30% yearly growth rate until 2022 and beyond
  - „Silicon Photonics – The key to data center connectivity“, R. Blum, Intel 2017
- **IBM Silicon Photonics Technology** – IBM T.J. Watson Research Center, 2008
  - 0.35 GFlops/W with „Roadrunner“ - World #1 in 2008
  - Rack-to-rack optics at some dollars per Gbps wo/ Silicon Photonics
  - Optical backplane at < 1 $/Gbps @ 10.000 ch @ 10 Gb/s/ch @ 10 mW/Gb/s/ch ► Reliability is an issue
  - Chip-to-chip optics at < 0.1 $/Gbps @ 100.000 ch @ 1Tb/s/ch @ < 10 mW/Gb/s/ch ► Reliability is a bigger issue
  - On-chip optical interconnects << 0.01 $/Gbps @ 100.000 ch @ 1Tb/s/ch @ < 1 mW/Gb/s/ch ►Reliability is a much bigger issue ► Forecast: 10 TFLOP on a 3D-chip @ 22nm
- **IBM Optical Interconnect Technology** – IBM Zürich Research Laboratory, 2005
  - Electrical disadvantage is so that „fast & dense“ and „fast & long“ is an issue
  - Optical Propagation properties nearly independent of modulation frequency ►carrier frequency 100.000 times higher
  - Benefits in bandwith-length product, innovative compute concepts ►fast, passive signal splitters
  - >10.000 channels on MCM @ distance 5mm-10cm
- **DARPA** – Ravi Athale, DARPA, June 2003
  - Optical on-chip interconnects program ►high channel density & speed, low power & coupling ►160 ch/cm @ 1mW/Gbps @ 10^{-15} BER
- **Standford University** – D. Miller, Stanford University, 2001
  - CMOS Scaling effects make optical interconnects desirable
  - Receiverless detection schemes for optical clocking
  - Advanced materials for detectors ►5fF @ 30x30µm



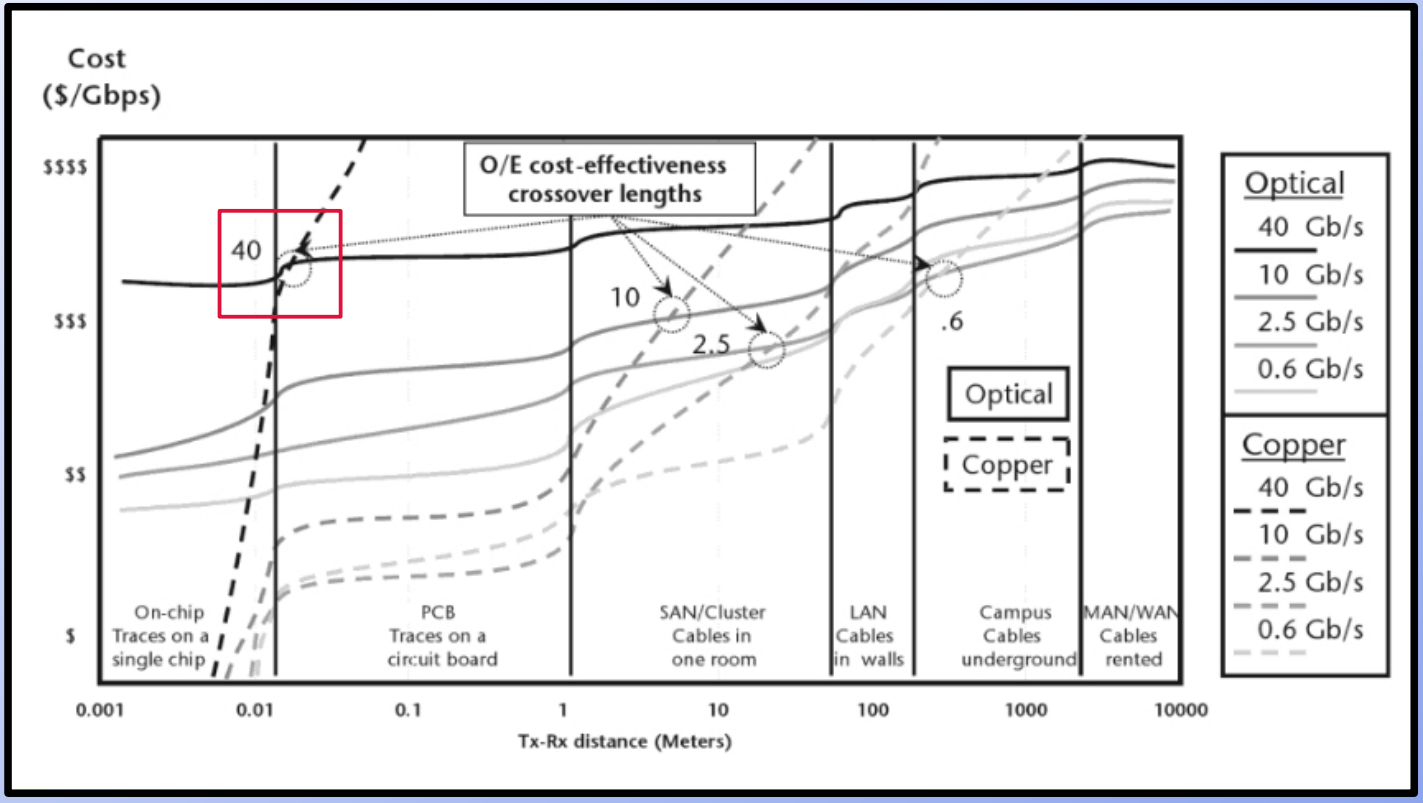
Active Optical Cable 40pJ/bit (Silicon Photonics).

On-chip optical interconnects (Silicon Nano-Photonics).



STI Cell chip with 256 Gflop (IBM Böblingen, 2006).

Link Cost vs. Distance & Bandwidth[1]



[1] M. Bahir, J. Meindl „Integrated Interconnect Technologies for 3D Nanoelectronic Systems“, ISBN-9781596932470, Artech House, 2008

Cu vs. Photonics vs. Nano-Photonics

Copper

On-chip & board-to-board

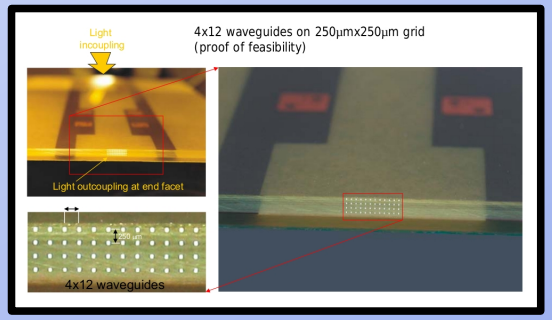
- Does not scale
  - RC-delay
  - Bandwidth
  - Energy/bit
- **Driver: power*cost/bit**
- 1000 Tx/chip @ 10Gbps
- High-perf CMOS used
  - ► BGA* Pitch scaling
  - ► Channel density

To-Off-Chip bottleneck![1]

Silicon Photonics

Rack-to-rack & board-to-board[2]

- Replace copper with fibre
- **Driver: cost/bit**
- 10 Tx/chip @ >10Gbps[1] aggregate
- <10 mW/Gbps/link @ <1$/Gpbs[2]
- High-perf CMOS not needed[1]
  - ► 130nm technology preferred[1]

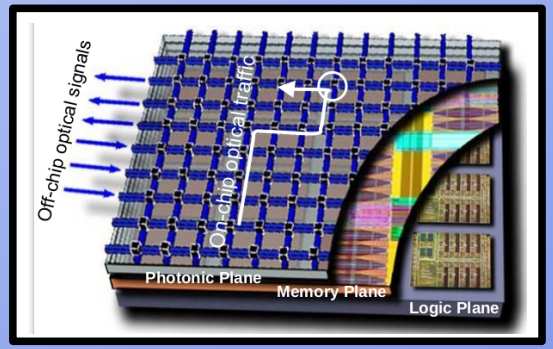


IBM Zürich Research Laboratory (ZRL),
R. Dangel, F. Horst, B. Offrein, 2008

Silicon Nano-Photonics[1]

Chip-to-chip & On-chip

- Replace copper with silicon
- **Driver: power/bit**
- 1000 Tx/chip @ 1Tbps[1] aggregate
- <1 mW/Gpbs/link @ <0.01$/Gbps[2]
- High-performance CMOS needed[1]
- 0.1cm-0.3cm distance
- 90nm process preferred (IBM 2020)



IBM: 300 cores/chip using 3D integration
https://newatlas.com/ibm-silicon-nanophotonics/25446/

[1] Y. Vlasov, IBM Thomas J. Watson Research Center, Yorktown, 2008
[2] IBM Terabus Project ► 240Gbps bidirectional @ 9 mW/Gb/s per unidirectional link
* BGA – Ball Grid Array = defines the distance of electrical connections on a silicon die

On-Chip Data Path Analysis

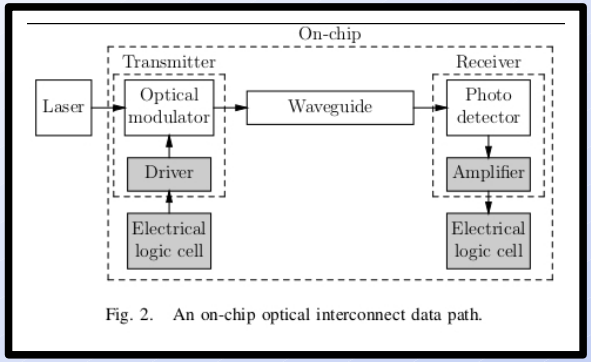


Fig. 2. An on-chip optical interconnect data path.

TABLE II

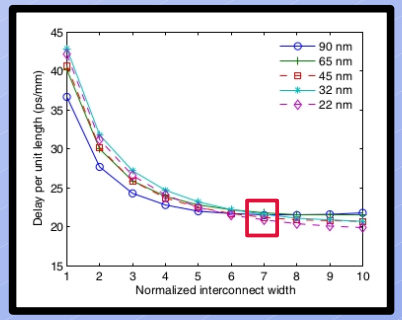
DELAY (ps) OF ELECTRICAL AND OPTICAL INTERCONNECTS.

Technology node	90 nm	65 nm	45 nm	32 nm	22 nm
Electrical	311.9	313.2	291.3	312.0	317.8
Optical	238.9	173.3	145.4	127.7	114.9

TABLE III

POWER (mW) OF OPTICAL AND ELECTRICAL INTERCONNECTS.

Technology node	90 nm	65 nm	45 nm	32 nm	22 nm
Transmitter	0.9	1.9	3.4	5.9	11.2
Receiver	0.5	0.5	0.3	0.3	0.3
Total optical	1.4	2.4	3.7	6.2	11.5
Electrical	9.8	16.9	21.7	33.4	45.3



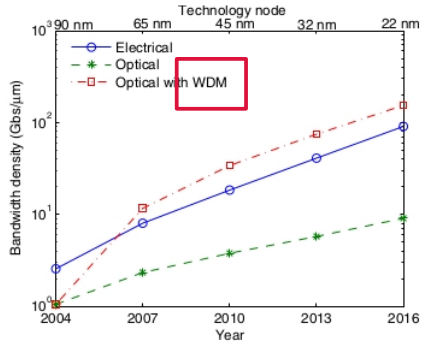


Fig. 5. Bandwidth density of electrical and optical interconnects.

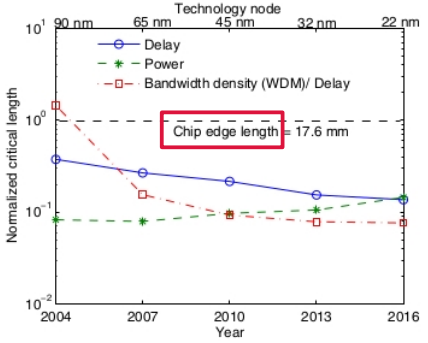


Fig. 6. Normalized critical length beyond which optical interconnect is advantageous over electrical interconnect.

- On-chip vs. Off-chip lasers
  - No Silicon lasers ►Opt. Modulator
- Power & Delay
  - Clock-frequency and leakage ►Tx power increases[1]
  - El. interconnect RC-delay increases exponentially[1]
- Bandwidth Density in Gb/s/µm
  - Single Mode Fibre ► >80THz
  - Dense Wavelength Division Mux (DWDM)

[1] Chen et. Al „Electrical and Optical On-Chip Interconnects in scaled microprocessors“, ISCAS 2005

Power Consumption

$$\frac{Energy}{Cycle} = \frac{Energy}{Instruction} \cdot \frac{Instruction}{Cycle}$$

$$P = \alpha \cdot f \cdot C \cdot U^2$$

$\alpha = switching\ factor\ [\%]$

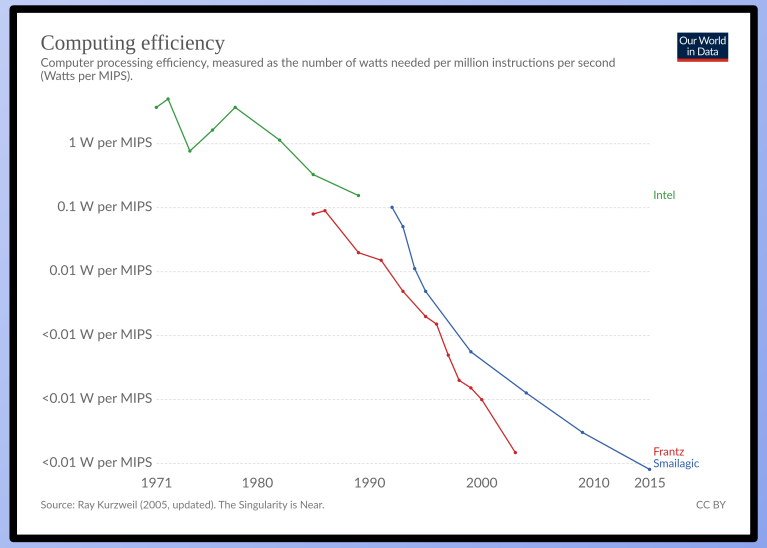
$f = frequency\ [Hz]$

$C = capacitance\ [F]$

$U = voltage\ [V]$

Power Consumption limits Compute Performance

- „Instruction per Cycle“
  - >1 for Superscalar Processors ► IBM P9: 128 @ 64 SMT-4
  - Energy consumption to be considered
- Computing-Power is limited by Power Density
  - Minimal „Energy per Cycle“ is design goal



Watts per MIPS over timescale (Source: Our World in Data)

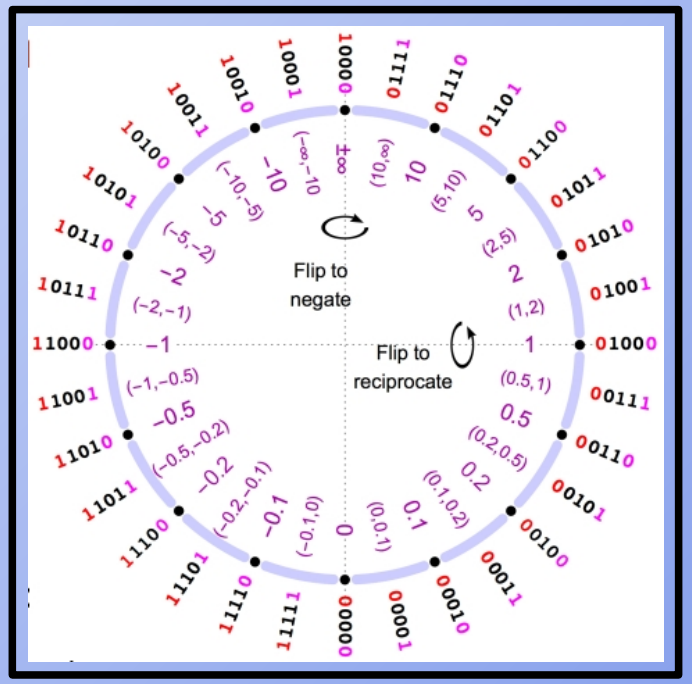
Reduce Power at Register Transfer Level (RTL) stage of a design[1]!

[1] A. Mutschler, Reducing Power At RTL Dependencies, February 13th, 2020, published on https://semiengineering.com

POSITs – A New FP-format

- Higher Dynamic Range $\log_{10}\left(\frac{W_{max,\,pos}}{W_{min,\,pos}}\right)$
- Decimals accuracy ► $10_{10} = 01110_{POSIT} \quad 0.1_{10} = 00010_{POSIT}$
- Lower RMS error[2] ► higher overall accuracy[2]
- Reproducable results across all architectures[2]
- Efficient hardware implementation[1]
  - Dot-product 3x to 6x faster than IEEE-754[1]

  ►Reduced silicon cost, energy & latency[1]
- Negate ► Two‘s Complement w/ first bit
- Invert ► Two‘s Complement wo/ first bit



4-Bit POSIT example [2].

[1] D. Biancolin, J. Koenig, “Hardware Accelerator for Exact Dot Product”, ASPIRE Laboratory, UC Berkeley
[2] J. Gustavson, „A Radical Approach to Computation with Real Numbers“, Supercomputing Frontiers and Innovations, 2016, Vol. 3, No. 2, DOI: 10.14529/jsfi160203

POSITs

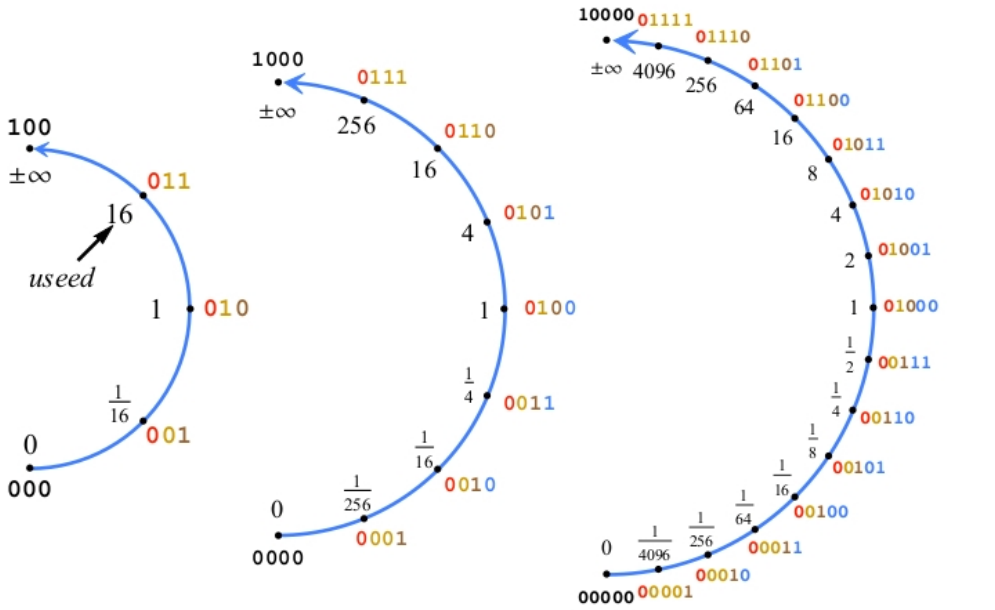


Figure 4. Posit construction with two exponent bits, $es = 2$, $useed = 2^{2^{es}} = 16$

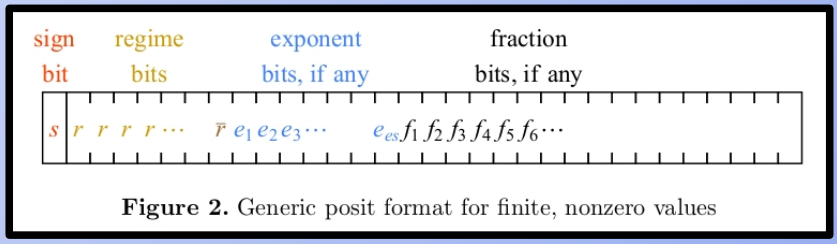


Figure 2. Generic posit format for finite, nonzero values

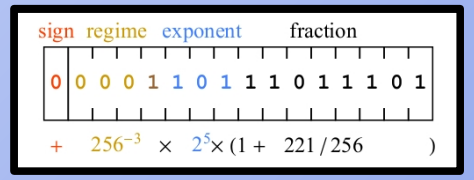
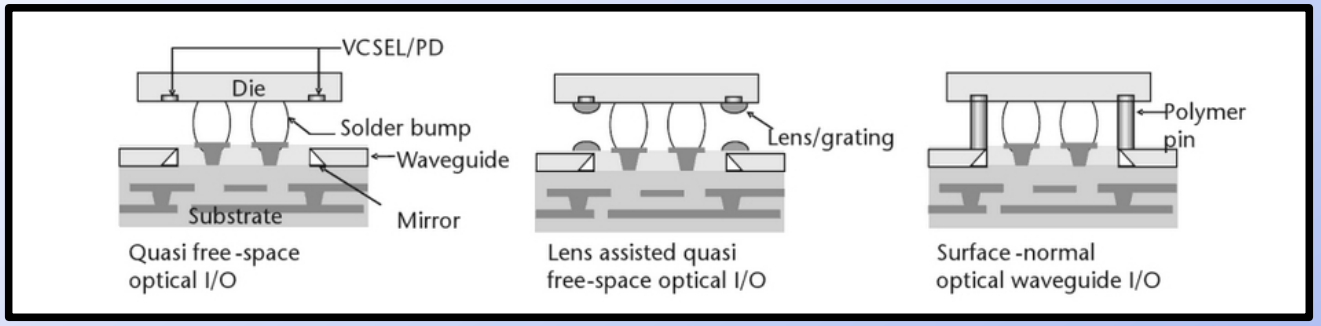
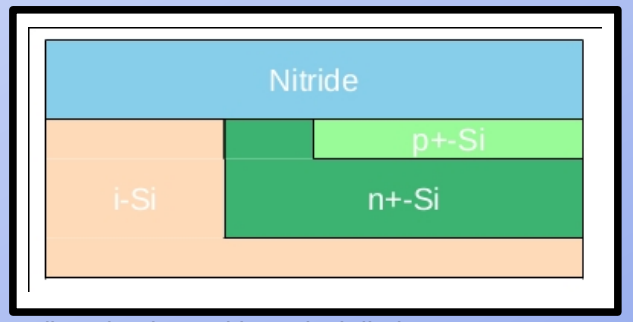


Table 3. Float and posit dynamic ranges for the same number of bits

Size, Bits	IEEE Float Exp. Size	Approx. IEEE Float Dynamic Range	Posit *es* Value	Approx. Posit Dynamic Range
16	5	6×10^{-8} to 7×10^{4}	1	4×10^{-9} to 3×10^{8}
32	8	1×10^{-45} to 3×10^{38}	3	6×10^{-73} to 2×10^{72}
64	11	5×10^{-324} to 2×10^{308}	4	2×10^{-299} to 4×10^{298}
128	15	6×10^{-4966} to 1×10^{4932}	7	1×10^{-4855} to 1×10^{4855}
256	19	2×10^{-78984} to 2×10^{78913}	10	2×10^{-78297} to 5×10^{78296}

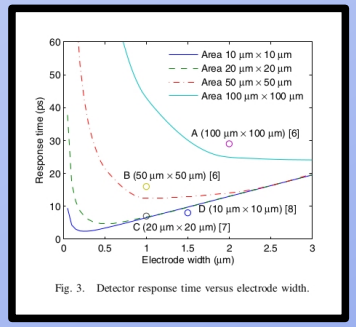
On-Chip Optical Detector[1]



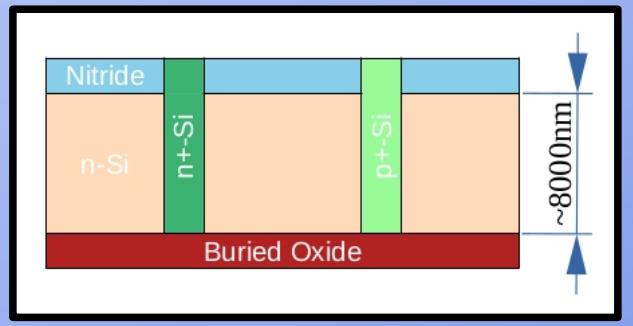


Bulk technology with vertical diode.

- **Free space** ► cost efficient
- **Lens assisted** free space & **Polymer pin** assisted
  ► tolerant to misalignments
- Response time ► detector capacitance
- Bulk technology ► Vertical design
- SOI technology ► Lateral design
  ► Compatibility to Standard CMOS SOI Technology



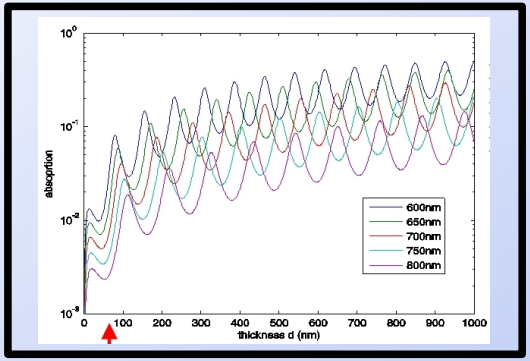
Response time over detector capacitance.[2]



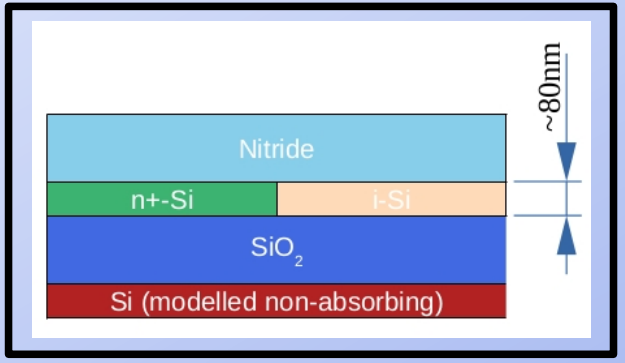
SOI technology with Deep Trench diode[1].
Deep-Trench normally used for embedded DRAM.

[1] Deep-Trench Diode (IBM T.J Watson Research Center, Yorktown) - Yang et. al. „High Speed Silicon Lateral Trench Detector on SOI substrate“ IEEE JLT, 2001
[2] Chen et. al., „On-Chip Copper-Based vs. Optical Interconnects: Delay Uncertainty, Latency, Power, and Bandwidth Density Comparative Predictions, 2006 International Interconnect Technology Conference

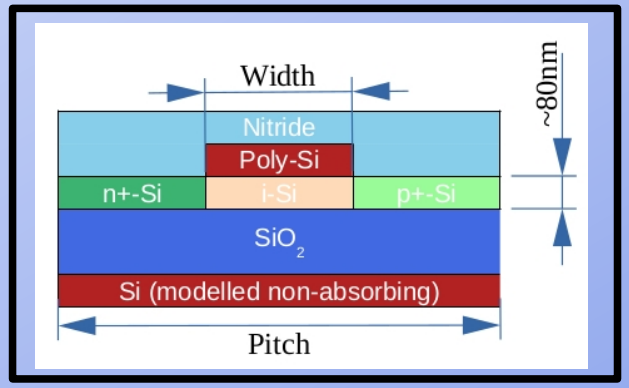
Optical Grating Detector



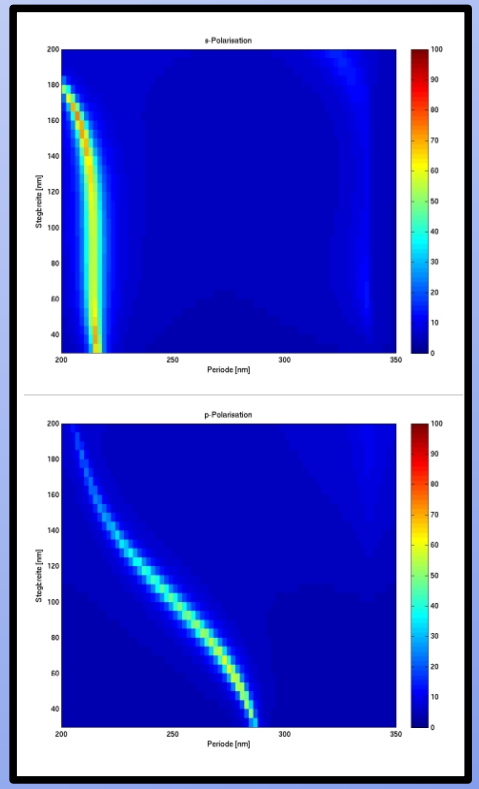
- Lateral SOI detector
  - Very low absorption
  - Resonant Fabry-Perot element
  - First ripple at 100nm
- Resonant grating coupler
  - Transistor gate forms grating
  - Standard SOI compatible



SOI technology w/ lateral diode.



SOI technology w/ lateral diode and resonant grating.



Simulation for TE-/TM-mode of a resonant grating with Px = Py (Brenner et. al., 2007)

[1] M. Fertig et. al., „Simulation of a 2d-grating assisted optical detector“, IBM report of Research & Development, 2007

Trench vs. Lateral Detector Analytics

- Quantum Effiency η & Responsitivity R

$$R = \eta \cdot \lambda \cdot \frac{q_e}{h \cdot c} \left[\frac{A}{W}\right]$$

- Built-in voltage (depletion region)

$$V_{bi} = \frac{K_{300}}{q_e} \cdot \log\left(\frac{N_A \cdot N_D}{N_{i,Si\,300}^2}\right) \quad [V]$$

- Thickness of depletion region

$$d_{depl}(V_{ext}) = \sqrt{\frac{2 \cdot \epsilon_{Si} \cdot \mu_0}{q_e} \cdot \frac{N_A + N_D}{N_A \cdot N_D} \cdot V_{bi}} \cdot \sqrt{1+\frac{V_{ext}}{V_{bi}}} \quad [m]$$

- Capacitance of depletion region

$$C_{depl}(V_{ext}) = \epsilon_{Si} \cdot \mu_0 \cdot \frac{A_{depl}}{d_{depl}(V_{ext})} \quad [F]$$

- Absorption length x of Silicon for $I(x) = I_0 \cdot e^{-\kappa k_0 x} = \frac{1}{e}$

$$x_\lambda = \frac{1}{\kappa \cdot k_0} = \frac{\lambda}{2\pi \cdot \kappa(\lambda)} \quad [m]$$

$x_{850\,nm} \approx 26\,\mu m$ $x_{670\,nm} \approx 5\,\mu m$

CMOS SOI process parameters

$$N_{i,Si\,300} = 1.45 \cdot 10^{10} \frac{1}{cm^3} \quad intrinsic\ carrier\ density$$

$$N_D = 1 \cdot 10^{20} \frac{1}{cm^3} \quad n-doping$$

$$N_A = 8.13 \cdot 10^{14} \frac{1}{cm^3} \quad p-doping$$

IBM T.J. Watson
„Deep Trench Detector“[1]

$$R = 0.45 \frac{A}{W}$$

$$\frac{d_{depl}}{Pitch} \approx 80\,\%$$

$$C \approx 450\,fF$$

$$\frac{I(8\,\mu m)}{I_0} \approx 30\,\%$$

$$\eta \approx 0.55$$

Deep-Trench vs. Lateral

$$\frac{Quantum\ Efficiency}{Capacitance} = \frac{\eta}{C}$$

Deep Trench detector[2]:

$$\frac{\eta}{C} \approx 1.2 \quad (Deep\ Trench)$$

Lateral detector[2]:

$$\frac{\eta}{C} \approx 1.4 \quad (Lateral)$$

[1] Yang et. al. „High Speed Silicon Lateral Trench Detector on SOI substrate“ IEEE JLT, 2001
[2] Source: P. Fischer, Heidelberg University, 2007

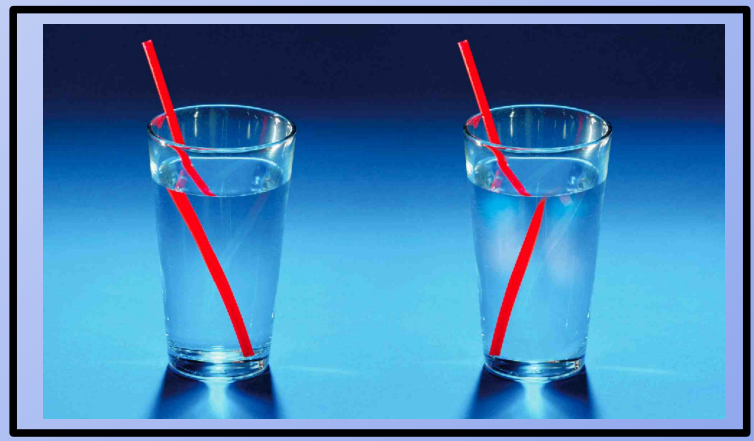
Negative Index Materials[1]

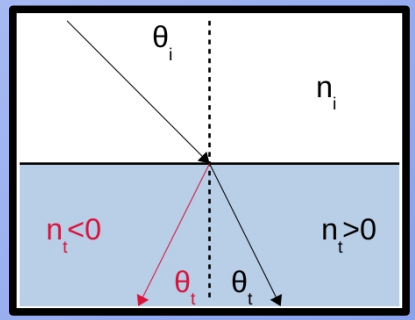
- Introduced by Veselago[1] (1968)
- Pioneering work by Smith[2] et. al. (2000)
- Electrodynamics of materials with negative ε and μ

$$\nabla\times\vec{E}=-\frac{1}{c}\cdot\frac{\partial\vec{B}}{\partial t} \qquad \nabla\times\vec{B}=\frac{1}{c}\cdot\frac{\partial\vec{D}}{\partial t} \quad \rightarrow \; Plane \;\; Waves \; \rightarrow \quad k\cdot\vec{E}=\mu\cdot\frac{\omega}{c}\cdot\vec{H} \qquad k\cdot\vec{H}=-\epsilon\cdot\frac{\omega}{c}\cdot\vec{E}$$

- Righ-Hand coordinate system (RHS) for positive ε,μ
  Left-Hand coordinate system (LHS) for negative ε,μ
- Snellius' Law of refraction $n_i\cdot\sin\theta_i=n_t\cdot\sin\theta_t$ with $n_t=\pm\sqrt{\epsilon\mu}$
- Plane Wave Propagation $\vec{E}(\vec{k},\vec{x})=\vec{E}_0\cdot e^{j(\vec{k}\vec{x}+\omega t)}$, $\vec{k}=n(\vec{x})\cdot k_0$
- Frequency dispersion $\epsilon(\omega),\mu(\omega)\rightarrow n(\omega)$ prevents „negative energy"

$$W=\epsilon E_0^2+\mu H_0^2 \quad (negative) \quad \rightarrow \quad \frac{\partial(\omega\epsilon)}{\partial\omega}E_0^2+\frac{\partial(\omega\mu)}{\partial\omega}H_0^2 \quad (positive)$$



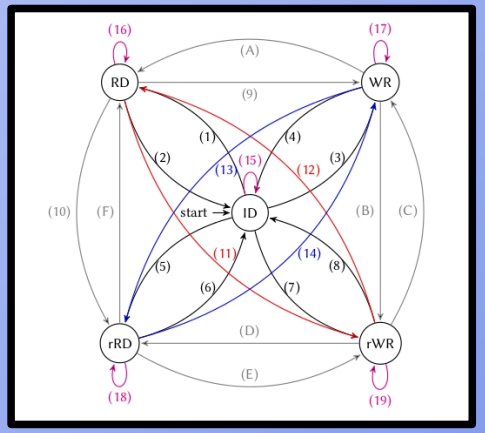


[1] Veselago, *Sov. Phys. Usp.* 10, 509 (1968)
[2] Smith, Padilla, Vier, Nemat-Nasser, Schultz, *Phys. Rev. Lett.* 84, 4184 (2000)

Universal **M**emory **A**utomaton

- Automated hardware implementation of complex algorithms
  - **Random Access Memory (RAM)** ►Cache Coherency Protocol
  - **Content Associative Memory (CAM)** ►Memory Management Unit
- Implementation of a **MESI Cache Coherency Protocol**[2]
  - Bachelor Thesis in 2018/19
  - Awarded the „MLP Leistungspreis“
- Hardware implementation of small accelerators
  - Perceptrons ► activation functions stored in memory
  - Near-Memory Multiply-Accumulate (MAC) units
  - Semi-custom design ► logic synthesis + full-custom memory
- Single-cycle memory operations supported only
  - ➔Muti-cycle state transfers ► **Multi-cycle Memory Access**
  - ➔Hierarchichal UMA designs ► **Arrays of Compute-Units**

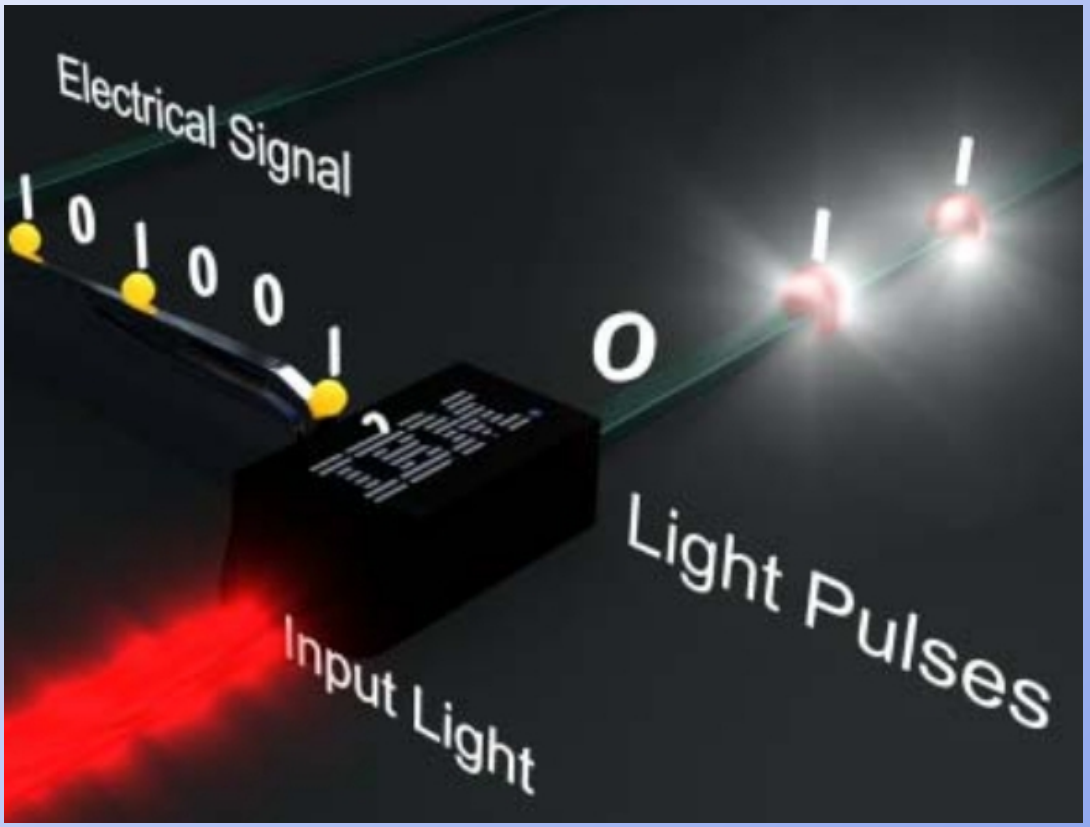
Expr	Condition
READ	res_n & cpu & rd
WRITE	res_n & cpu & wr
R_READ	res_n & !cpu & rd
R_WRITE	res_n & !cpu & wr
NOP	!res_n || (!rd & !wr)
TAG_MATCH	POP(R,addr(3:0))==addr(15:4))
IS_MODIFIED	POP(MESI,addr(3:0))==4'b1000)
IS_EXCLUSIVE	POP(MESI,addr(3:0))==4'b0100)
IS_SHARED	POP(MESI,addr(3:0))==4'b0010)
IS_INVALID	POP(MESI,addr(3:0))==4'b0001)
SET_TAG	PUSH(R,addr(3:0),addr(15:4))
SET_MODIFIED	PUSH(MESI,addr(3:0),4'b1000)
SET_EXCLUSIVE	PUSH(MESI,addr(3:0),4'b0100)
SET_SHARED	PUSH(MESI,addr(3:0),4'b0010)
SET_INVALID	PUSH(MESI,addr(3:0),4'b0001)



[1] M. Fertig „Universal Memory Automaton“, MPC Workshop Proceedings, University of Applied Sciences Mannheim, Feb 2020

[2] http://www-home.htwg-konstanz.de/~MFERTIG/files/tools/VERIGEN_FREE.tar

Modulator Principle of Operation



Quantum Computing

M. Bremner, Sidney University of Technology

The advance of Quantum Computing is not about speed but about complexity of algorithms.

C. Calude, Auckland University

When talking about Quantum Computing some classify conventional computing as obsolete – but this is not true. It is a competition that is not yet decided.

J. Preskill, California University of Technology (Caltech)

The correct answer about when the point of Quantum Supremacy is achieved depends on how fast conventional algorithms are.

R. Josza, University of Cambridge

We were not able to figure out why some algorithms simulate easy and others not on conventional systems. Some experts raised the question if lots of the supreme Quantum-Algorithms are in reality very normal.