FUTURE COMPUTER ARCHITECTURES ACTIVITIES

Central Institute of Engineering, Electronics and Analytics – Electronic Systems (ZEA-2)

18-09-19 I STEFAN VAN WAASEN

OUTLINE

- Future computer archtectures
  - Why do we need them?
- Neuromorphic Computing – NC
  - State-of-the-art systems
  - Next generation challenges
  - FZJ approach and set-up
- Quantencomputing – QC
  - Today‘s Quantencomputer approaches
  - Vision of fully scalable Universial Qantencomputer
  - Development approach
    - High level system model
    - Bottom-up implementation

FUTURE COMPUTER ARCHITECTURES

Why do we need them? – Limitations of HPC

- Supercomputer (HPC)
  - Even the most powerful HPCs show clear limitations in solving …
    - … huge problems
      - Classical examples
        - Factorization of big numbers – cryptography, security
        - Search algorithms

    - … „brain-related“ topics
      - Classical examples
        - such as object recognition or prediction in natural environments
        - learning from few examples in new or changing environments
        - involving novelty and ambiguity

JÜLICH Forschungszentrum

FUTURE COMPUTER ARCHITECTURES

Why do we need them? – Possible improvement by beyond von-Neumann architectures

- Neuromorphic computing (NC) – brain-inspired computing architectures
  - Mimicking structure and dynamics of biological neuronal systems for “new” solutions
    - “Cognitive” computing applications (deep learning, robotic, autonomous driving, …)
    - Neuroscience applications
      - Simulation tool for brain “understanding” – Neuroscience research
- Quantencomputing (QC) – make use of quantum physics phenomena
  - In general no new type of problems can be solved, but …
  - … offers great potential to increase solvable complexity
    - Quantensimulations
    - Complex chemical/biological problems (pharamceuticals, material research, …)
    - Financial market research

JÜLICH Forschungszentrum

NEUROMORPHIC COMPUTING

State-of-the-art overview

(Nawrocki et al., 2016)

Name	From	Type of synapse	Number of synapses	Type of neuron	Number of neurons	Technology/material	Learning	Power consumption	Size/area	Demonstration/application	Notes
Neuromorphic retinas	INI			spiking photoreceptor circuit	16x10^3	350 nm CMOS	no	23 mW	36 mm^2	high (10k fps) speed video imaging	available commercially
Neurogrid	Stanford	spiking (FPGA/RAM)	6x10^9	I&F	10^6	250 cm CMOS	STDP	5 W	168 mm^2	affordable neuromorphic supercomputer for biologic simulations	mixed analog/digital multichip system
ROLLS	INI	spiking (CMOS)	128,000	I&F	256	180 nm CMOS	spike-based plasticity/STDP/WTA	4 mW	51.4 mm^2	cortical-like computational modules	12.2x10^6 TFTs; on-chip configuration of network connections
SpiNNaker	University of Manchester	spiking (emulated)	5x10^7	spiking	2x10^4 (goal: 10^9)	180 nm ARM processors/software hybrid	configurable; spike-based plasticity	100 nJ/neuron + 43 nJ/synapse	65,536, 18-core ARM processors	learning temporal sequences of neural activity	neuromorphic architecture for simulating spiking neural networks
NPU	Qualcomm			spiking		20 nm ARM processor;	yes			image and sound processing	Snapdragon 820's Zeroth Neural Chip
TrueNorth	IBM	spiking (CMOS)	256x10x^6	I&F	10^6	28 nm CMOS	no	60 mW	4.3 cm^2	object recognition in live video	5.4x10^9 TFTs
NOMFET	University of Lille	spiking	single device		none	organic; pentacene + Au NPs	STDP			facilitating and depressing synaptic behavior	3-terminal organic memristive device
Organic memristor	University of Parma	memristive	single device		none	organic; PANI	yes			Boolean AND, OR, NOT gates	3-terminal organic memristive device
OECT	Saint-Etienne School of Mines	spiking	single device		none	organic; PEDOT:PSS	STDP		single device: 2.4 mm x 0.25 mm	future brain machine interface	demonstration of spiking activity in polymeric material
PNC	University of Denver	memristive	2/neuron	non-spiking	1	organic; PEDOT:PSS, PQT-12	no	7.5x10^-5 (P/IPS)	2 chips, 1'x1' each	linear classification for soft robot	organic hardware neural network
BrainscaleS	Heidelberg University	spiking, conductance based	10^4 / neuron	I&F (AdEx), multicompartment models	10^3-10^4 / wafer	wafer-scale analog VLSI, 180nm CMOS	STDP	10pJ / synaptic event	20 cm wafer, 20 wafers in 7x19"racks	neuroscience simulation platform, plasticity & learning	mixed analog/digital, acceleration factor 10,000

only few large-scale systems approaches

aiming at plastic natural-density connectivity

SpiNNaker – Manchester University

BrainScaleS – Heidelberg University)

JÜLICH Forschungszentrum

NEUROMORPHIC COMPUTING

Generation 1 and 1.5 of NC systems for Neuroscience research

SpiNNaker

- Fully digital ARM based architecture
- Highest flexibility
- Real-time „acceleration“: x 1
- Energy cons.: HPC-like

BrainScaleS

- Mixed-signal (analogue/digital)
- Low flexibility
- Real-time „acceleration“: x 1000
- Energy cons.: ?

JÜLICH Forschungszentrum

NEUROMORPHIC COMPUTING

Challenges for next generation systems

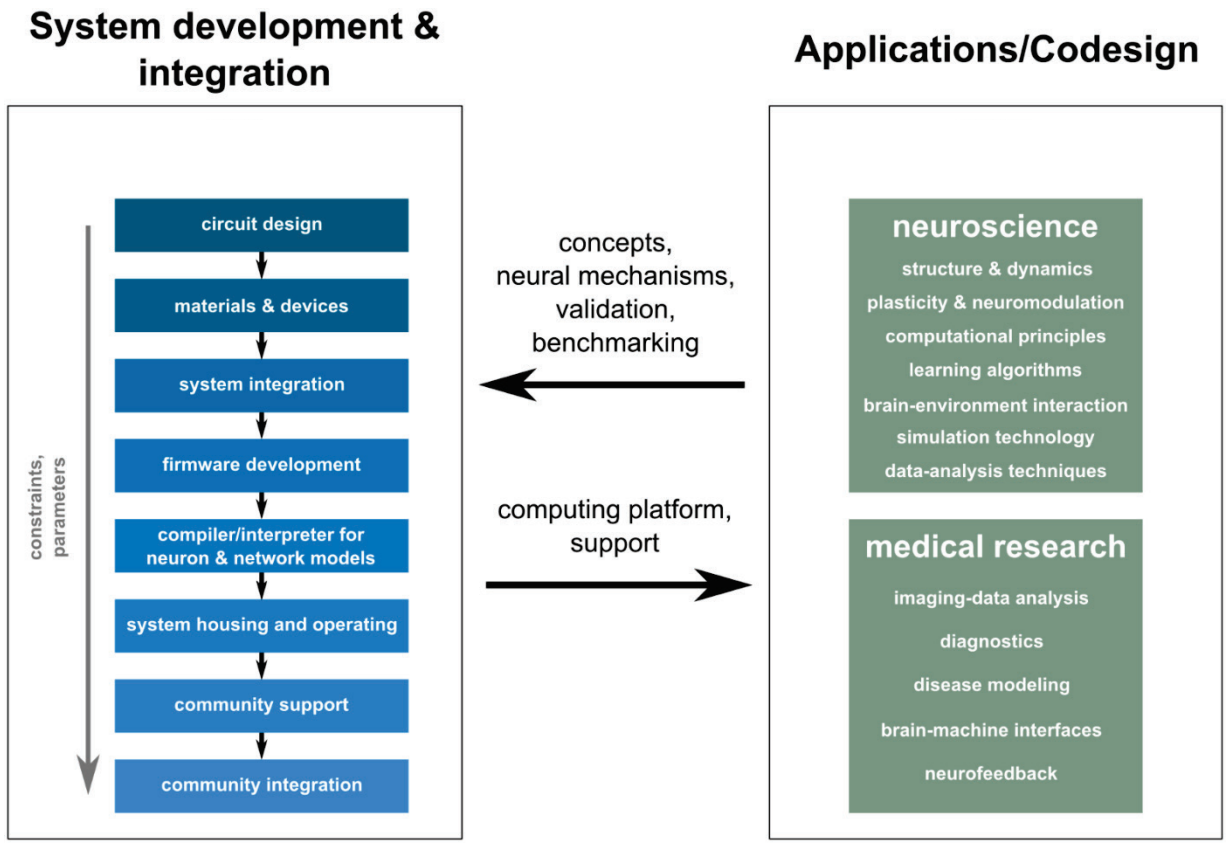
- Combine the best of both architectures on CMOS
  - Enable high connectivity level
    - Enabling of full 3D-integration
    - Make use of communication system principles (time, frequency, code multiplex)
      - Increase number of operating dimensions
  - Certain flexibility for en-/disabling „biological“ ingridients
    - Spike/non-spike principals, neuron/synapse heterogenity, switchable reproducability, …
- Early consideration of SW requirements
  - One of today‘s bottle necks: mapping of SW on HW
  - Early architecture considerations for optimized SW/HW interaction
    - Co-development with Neuroscience „users“
      - Early prototyping on flexible HW (e.g. FPGA)

JÜLICH Forschungszentrum

NEUROMORPHIC COMPUTING

Uniqueness capability of FZJ

- NC research is highly interdisciplinary and multi-level
  - Broad spectrum of competences and long-term expertise inside FZJ
    - Incl. ext. leading experts in NC systems (BrainScaleS, SpiNNaker)
      - rapid import of knowledge into FZJ
  - Unique capability and opportunity to develop a large-scale research & user facility – 10 year scale

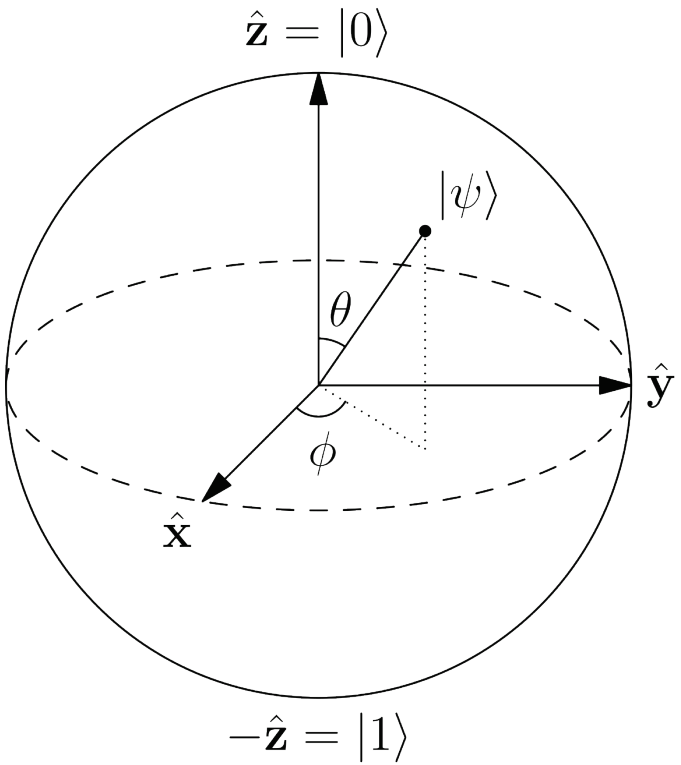


QUANTENCOMPUTING

Fundamental Element for QC – Quantum bit (Qubit)

- Two-state quantum-mechanical system (possible superposition of both states)
  - Simple example: vertical/horizontal polarization of single photons
- Representation by Bloch sphere
  - Basis states (vectors): $|0\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $|1\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

    with $|\psi\rangle = \alpha \cdot |0\rangle + \beta \cdot |1\rangle$ and $\alpha^2 + \beta^2 = 1$
  - *n* qubits require 2^n complex numbers for description
- Quantum Entanglement
  - Non-local property as expression of higher correlation
  - Unique resource for quantum computing



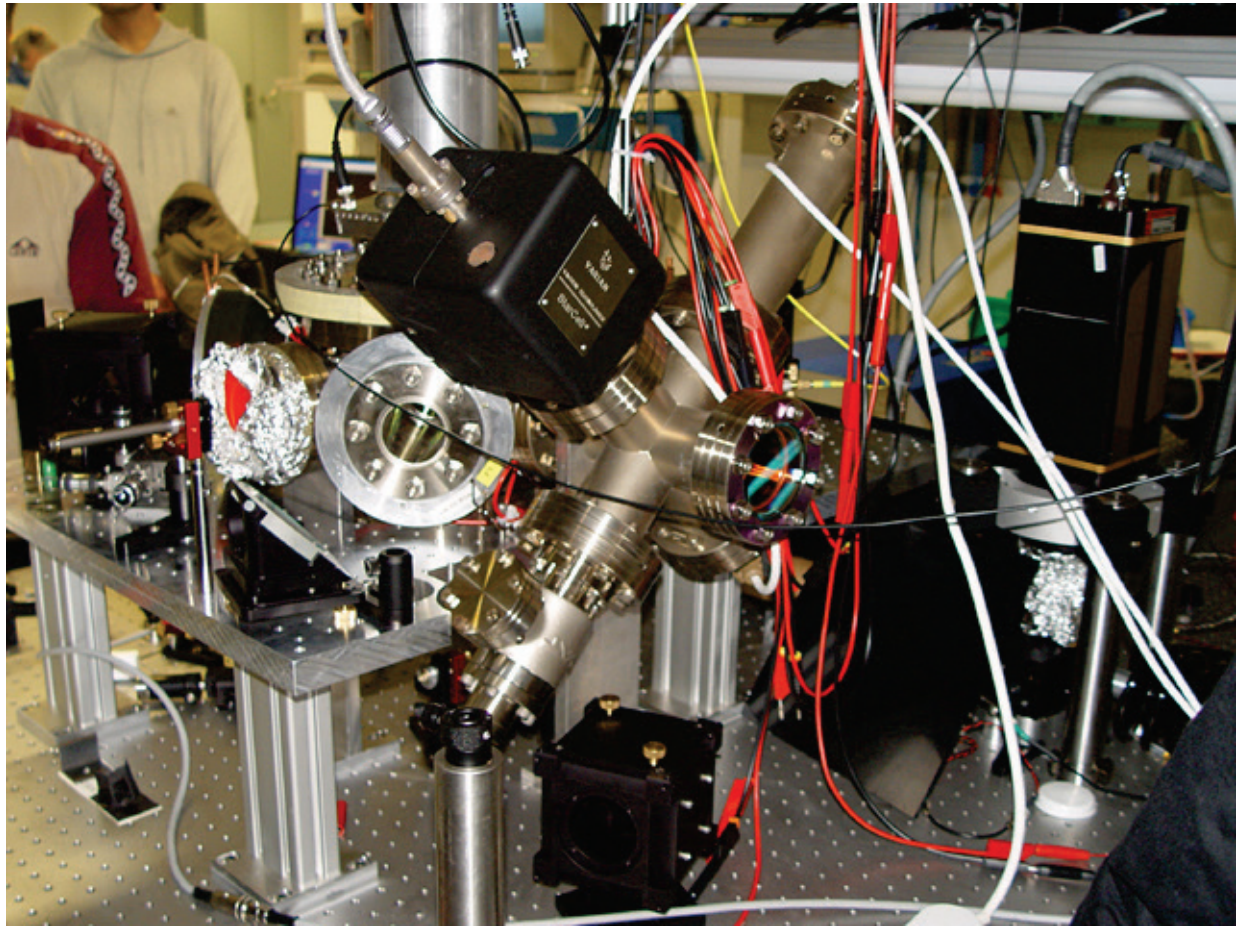
By Glosser.ca – Own work: Bloch sphere

JÜLICH Forschungszentrum

QUANTENCOMPUTING

Qubit – Some physical represantations

- Ions in ion-traps
  - Single ions captured (chain) in vacuum by electromagnetic fields
  - Electronic states (e.g ground state level – excited level)
  - Initialization by laser light, manipulation by magnetic transitions



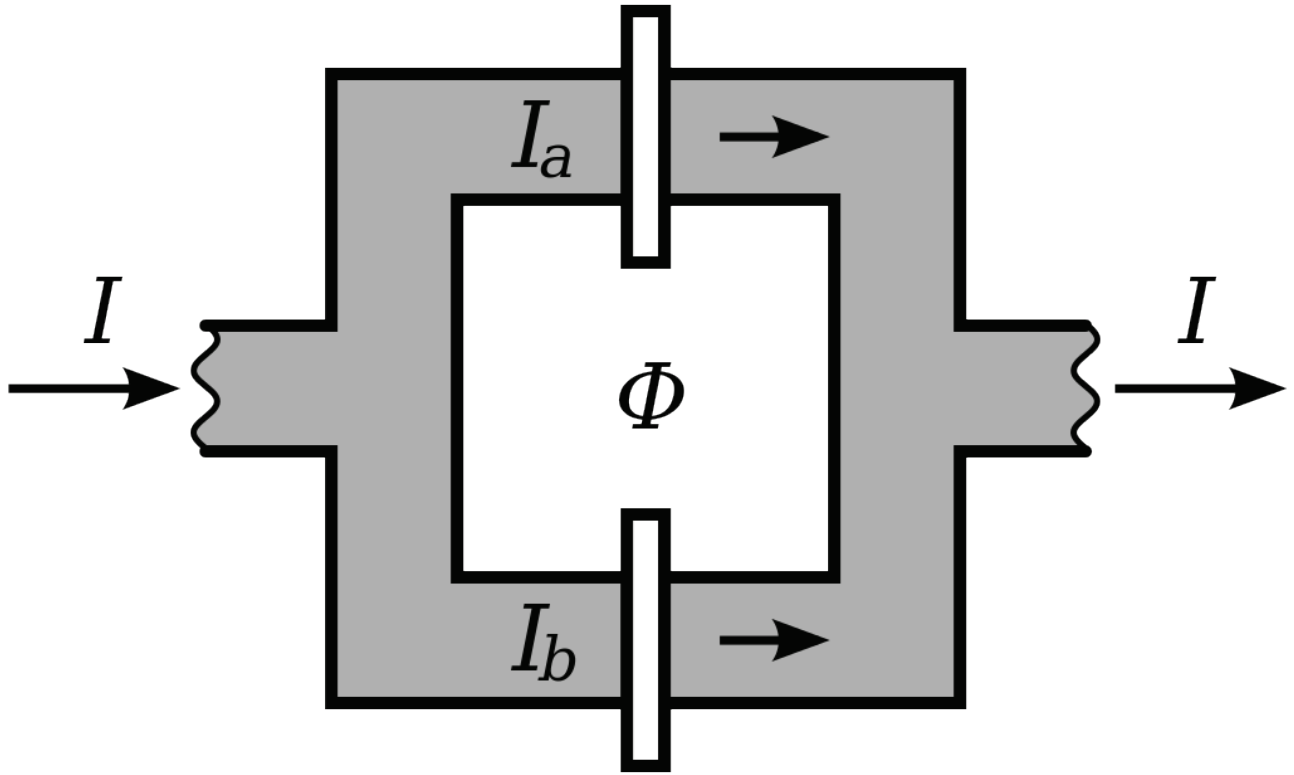
By Mnolf – Photo taken in Innsbruck, Austria

JÜLICH Forschungszentrum

QUANTENCOMPUTING

Qubit – Some physical represantations

- Ions in ion-traps
- Superconducting qubits by use of Josephson junctions
  - Seperated weakly correlated wave functions (quantum tunneling)
  - Different archetypes
    - Charge
    - Flux
    - Phase
  - Under research by Google, Microsoft, IBM, Intel

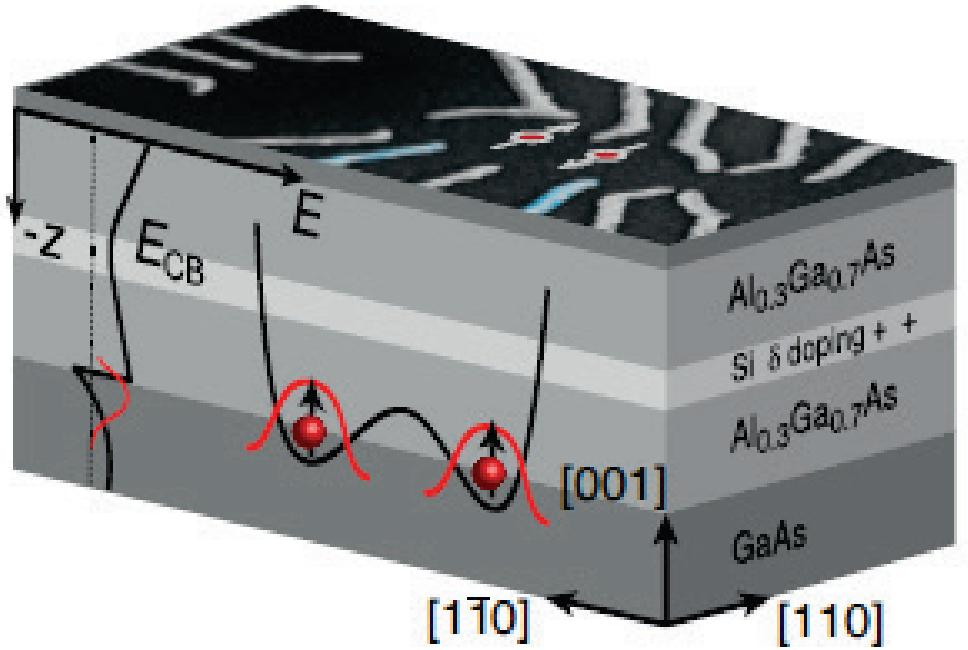


By Miraceti – Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=13302041

QUANTENCOMPUTING

Qubit – Some physical represantations

- Ions in ion-traps
- Superconducting qubits by use of Josephson junctions
- Electrons in quantum dots
  - Spin or charge types possible
  - Semiconductor based technology (GaAs, SiGe, Si-CMOS)
  - Promising towards scaling capabilities

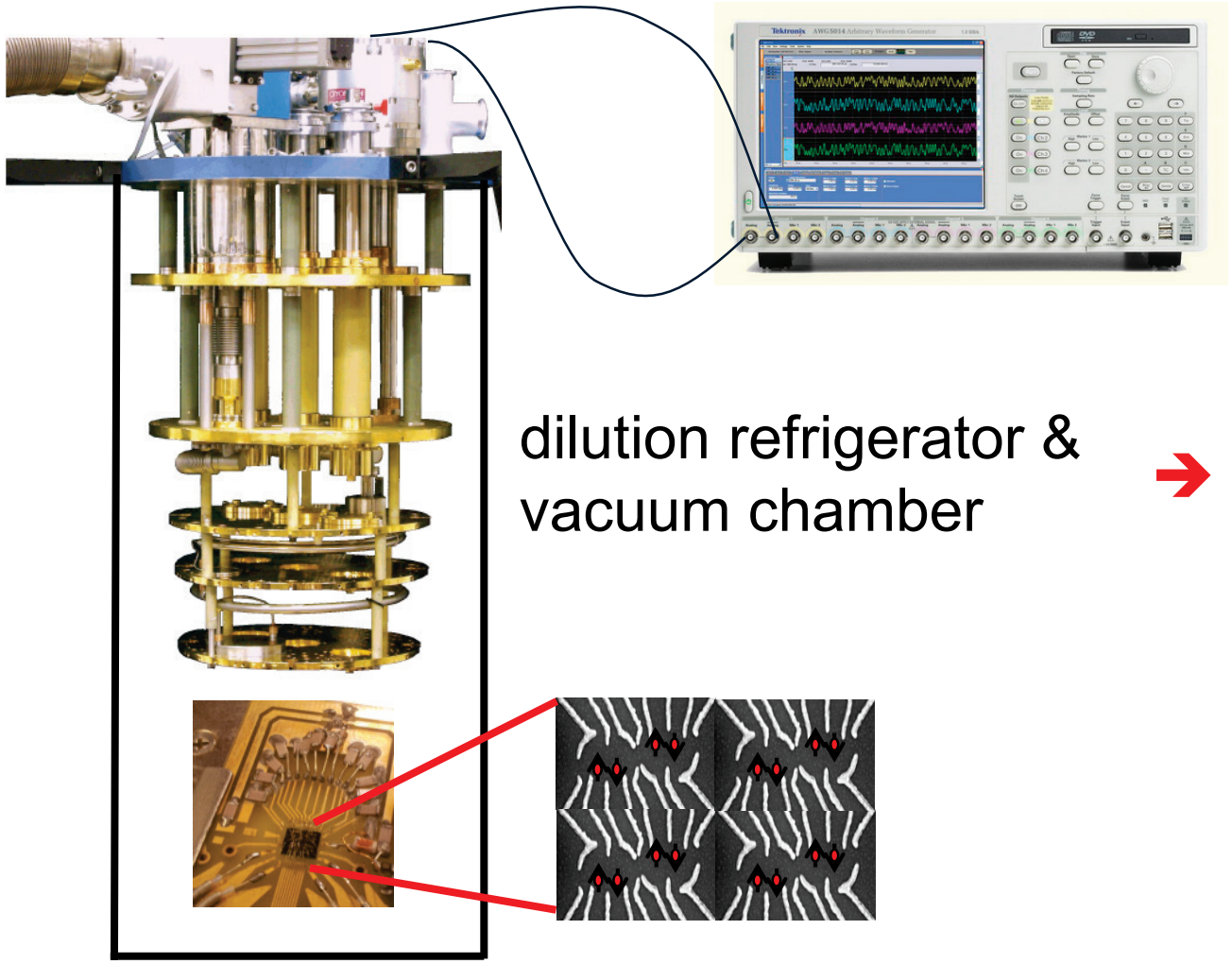


Tim Botzem – „Coherence and high fidelity control of two-electron spin qubits in GaAs quantum dots,“ PhD Thesis, p. 7, Figure 2.2.: Device Layout. Online: http://publications.rwth-aachen.de/record/689507, 14.08.2017

JÜLICH Forschungszentrum

QUANTENCOMPUTING

State-of-the-art measurement approach – Single or few qubits



room temperature

pulse generator – 1 for 2 qubits
- **min.** 2 coaxial cables per qubit

dilution refrigerator &
vacuum chamber

➔ availability of ≈1 mW cooling power

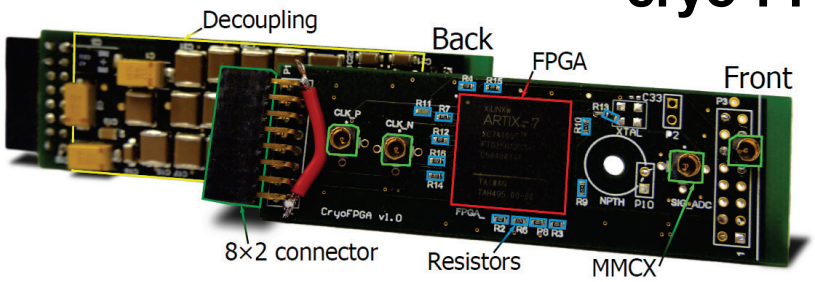
JÜLICH Forschungszentrum

QUANTENCOMPUTING

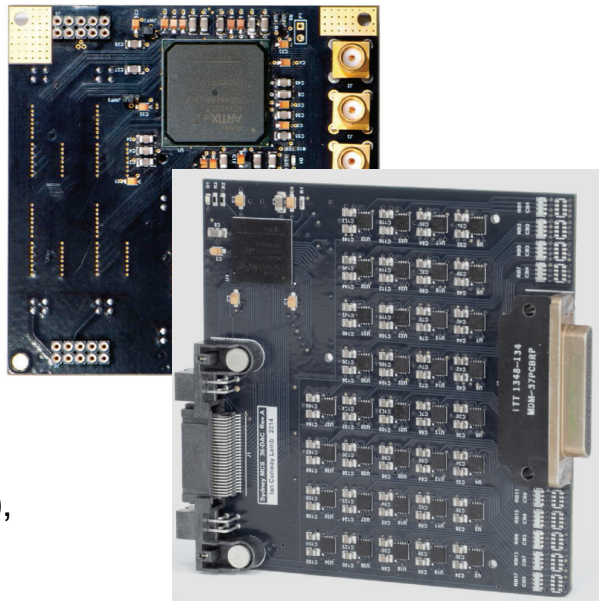
First “real“ QC approaches

- ‚Brute force‘ scaling for operation of around 100 qubits
- Further scaling clearly limited

cryo-FPGA

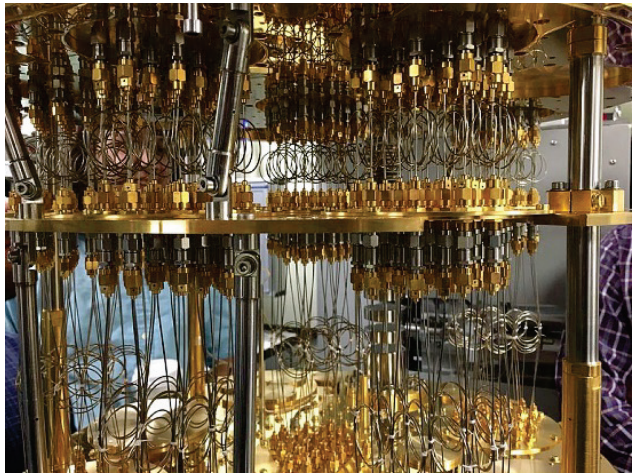


Homulle et al., IEEE Trans. Circuits & Systems 63 (2016), pp. 1854-1865



Lamb et al., Rev. Sci. Inst. 87 (2016), pp. 1-7

massive parallelization



IBM Research

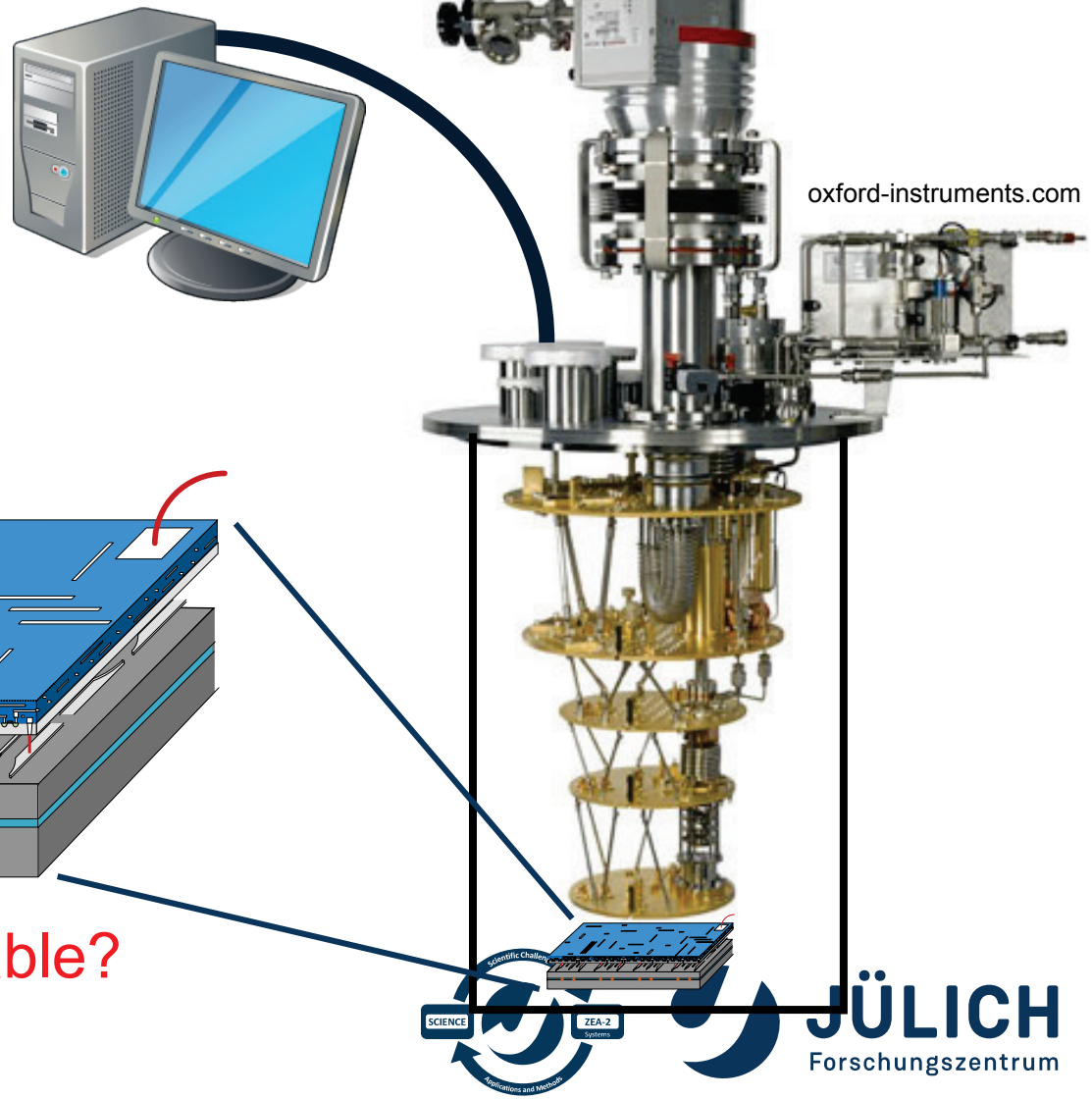
Mohseni et al., Nature 543 (2017), pp. 171-173

JÜLICH Forschungszentrum

QUANTENCOMPUTING

Vision of fully scalable „Universial“ QC – Fully integrated QC chip with simple interface

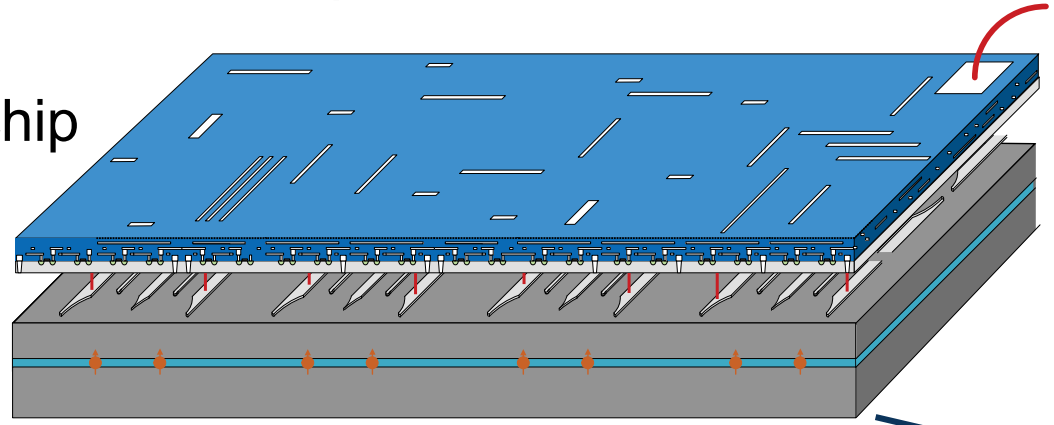
- Extremely-low-power control/read-out chip in „standard“ electronics → hypothesis: semi-conductor based solution required (e.g. CMOS)
- High density 3D interconnect
- Fully scalable Qubit device/chip



elp-control/read-out chip

3D interconnect

Qubit device/chip



➔ How the get control electronics fully scalable?

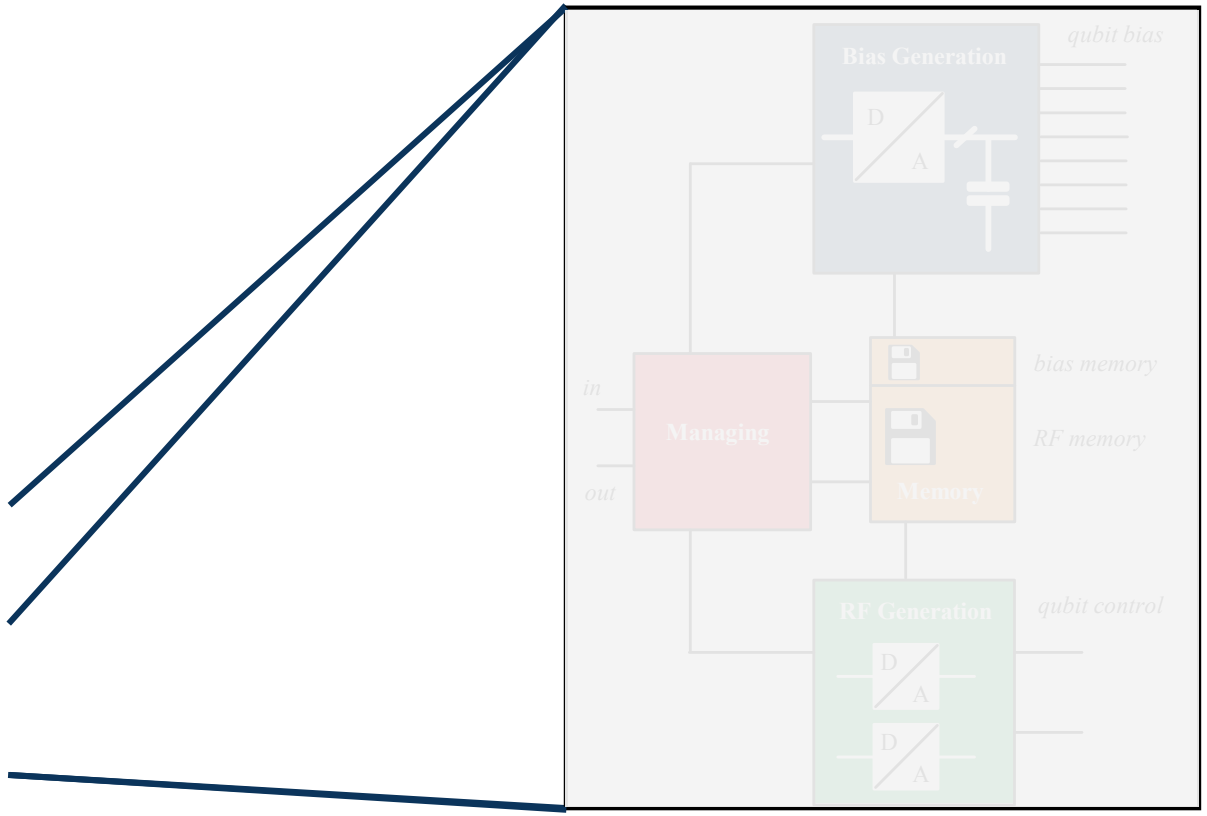
QUANTENCOMPUTING

Development approach – Which “interfaces” need to be considered?

Possible QC system layer model

- Quantum Algorithm → Quantum SW
- Translator/ Compiler
- Quantum SW → Operation
- Error Correction, Operation Execution
- HW Control/ Read-out
- Qubits

- Error Correction/Operation Execution: control information
- Qubits: configuration/controll and read-out



Possible control architecture

JÜLICH Forschungszentrum

QUANTENCOMPUTING

Development approach – Combine Top-Down and Bottom-Up design approach

Top-Down design

- Start from high level system model/ design
- Specify system and interfaces by model
- Elaborate and optimize system set-up by state-of-the-art implementations

Bottom-Up implementation

- Start from basic components
- Optimize performance trade-off
- Design and test component
- Assemble subsystem with several components for pre-system tests

Combine results from both approaches for functional system with detailed component characteristics

JÜLICH Forschungszentrum

QUANTENCOMPUTING

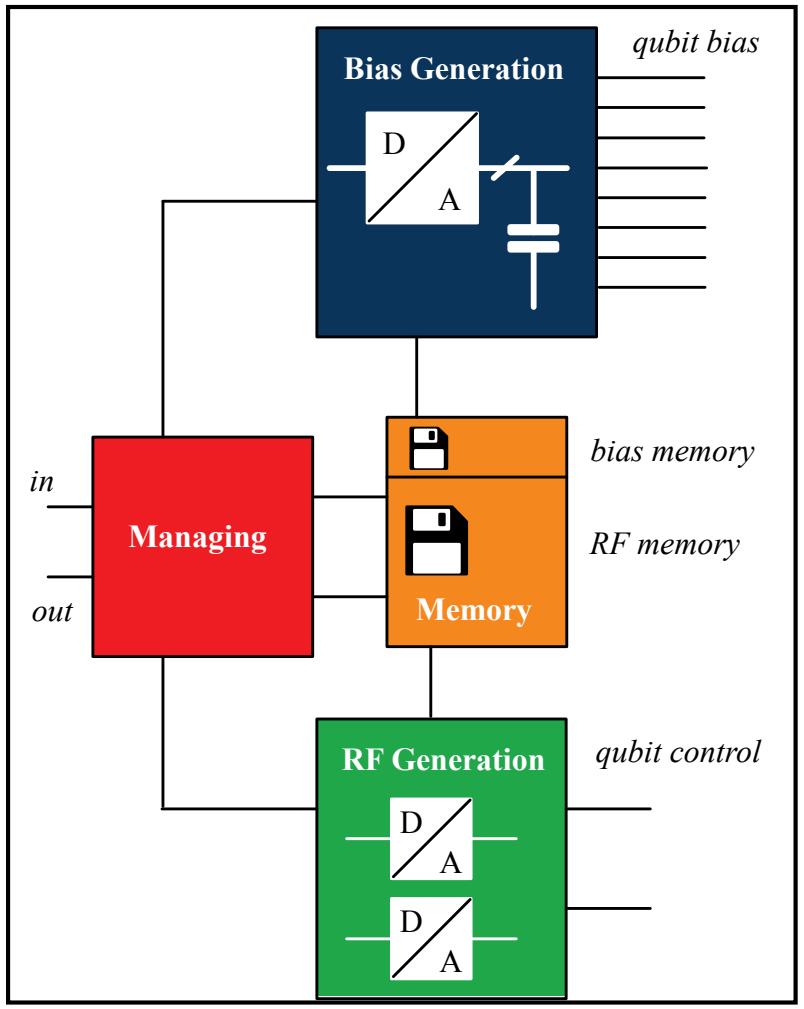
Requirements based on GaAs spin qubits

- Up to 8 uncorrelated bias voltages per Qubit, forming potential wells
- 2 pulse electrodes for Qubit operation
- Readout by reflectometry
- Key performance indicator: fidelity of Qubit Gates

Characteristic	Specification
DC voltage range	−1 V to 0 V
DC voltage stability	≲ 20 µV
DC Stepsize	250 µV (≙12 bit)
Pulse voltage range	± 4 mV
Pulse sampling rate	250 MHz
Overall cooling power budget @ 100 mK	1 mW

QUANTENCOMPUTING

High level system model – Functionality

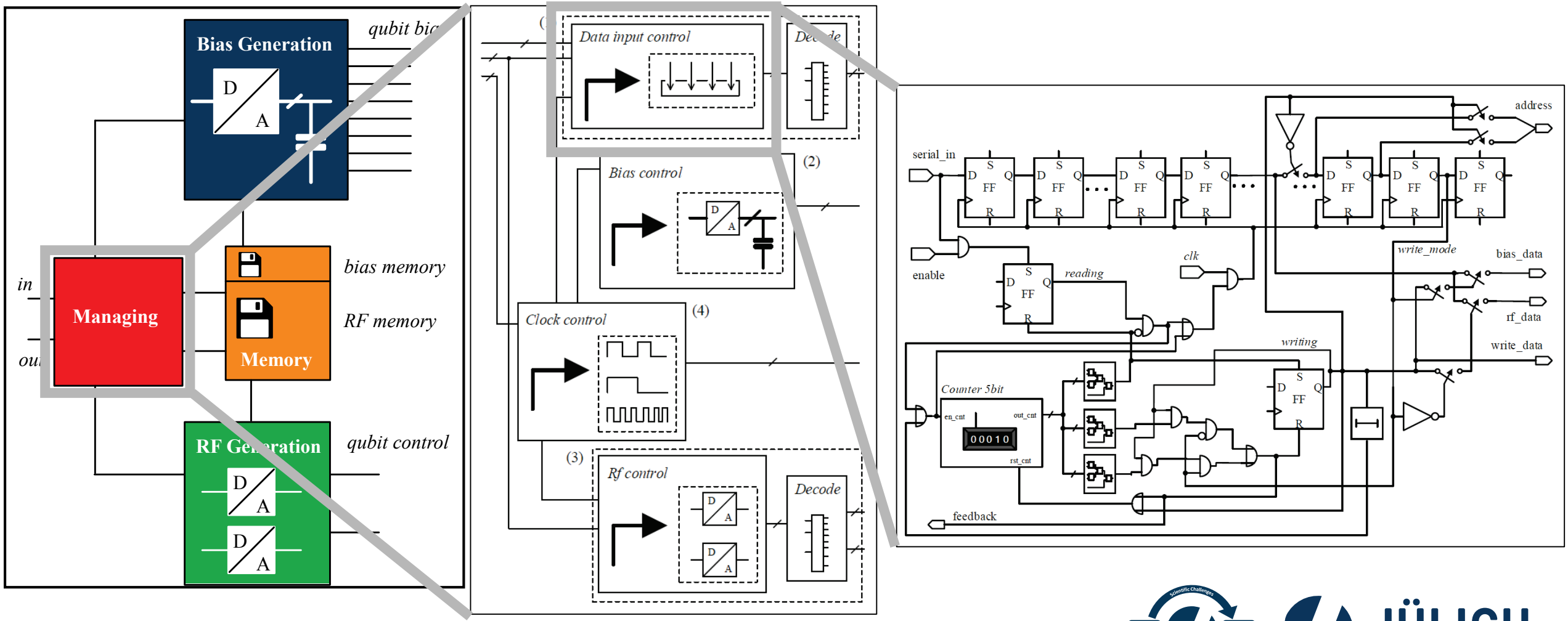


- Main functional components
  - Memory
    - Storage of voltage levels and pulse forms
  - Bias Generation
    - Generates DC voltages for confinement of electrons
  - RF generation
    - Generates pulses for operational control of Qubit
  - Managing
    - I/F to upper levels and sequencing

JÜLICH Forschungszentrum

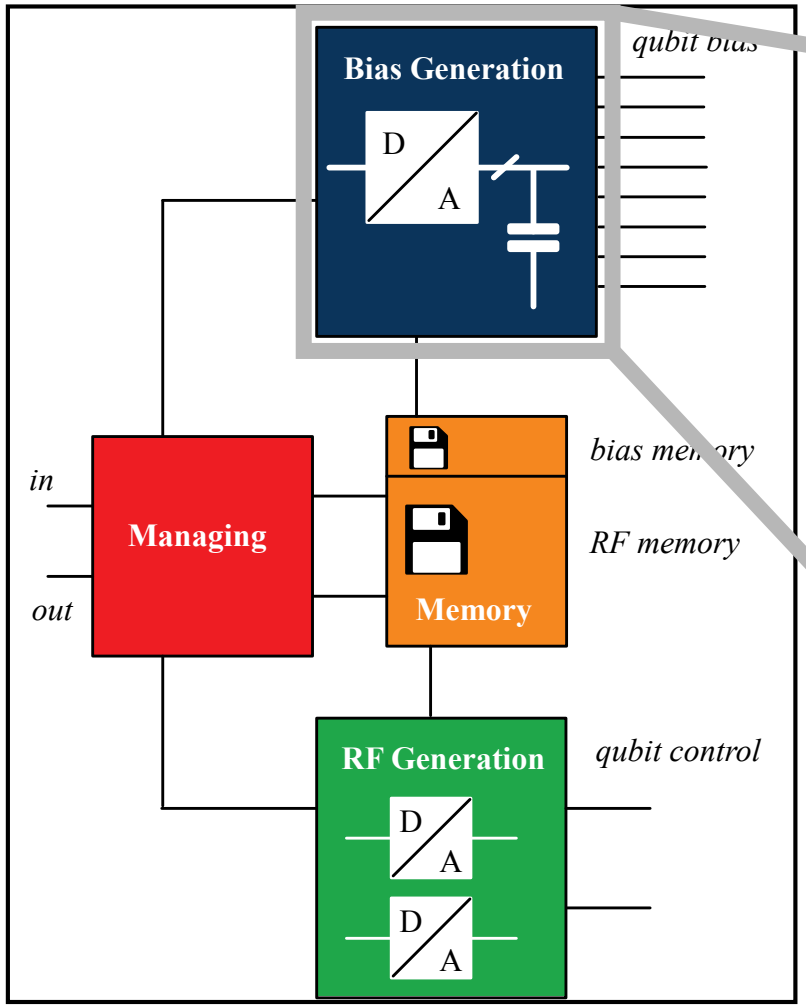
QUANTENCOMPUTING

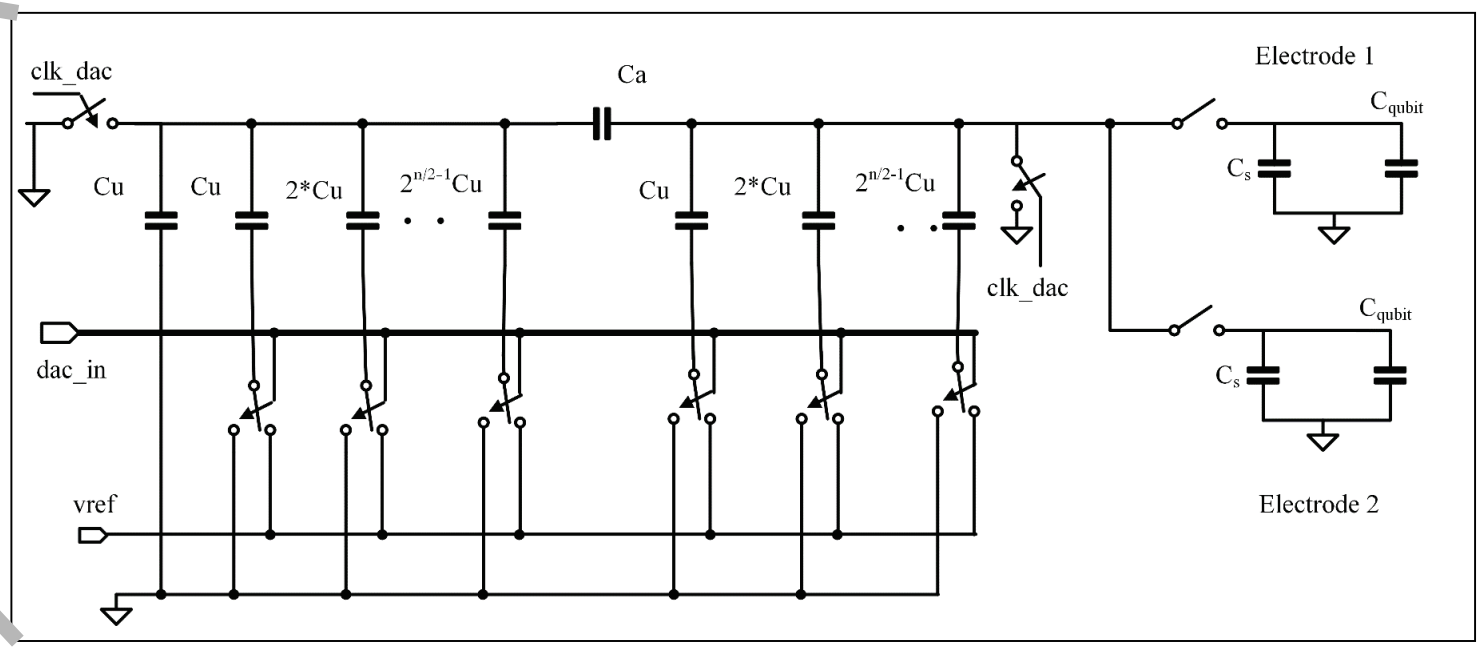
High level system model – Control block



QUANTENCOMPUTING

High level system model – Bias voltage generation



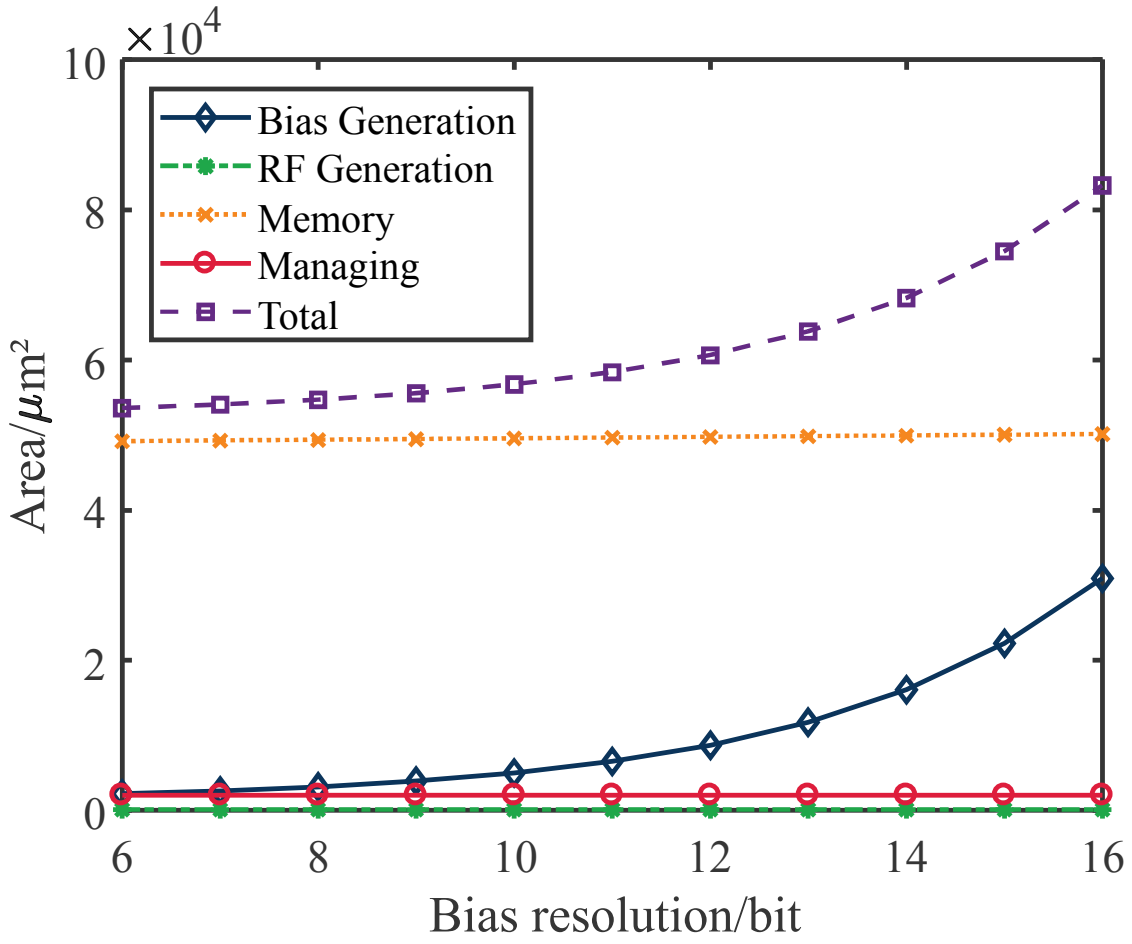


- Capacitor based Digital-to-Analog Converter (DAC)
- Multiplex and hold functionality

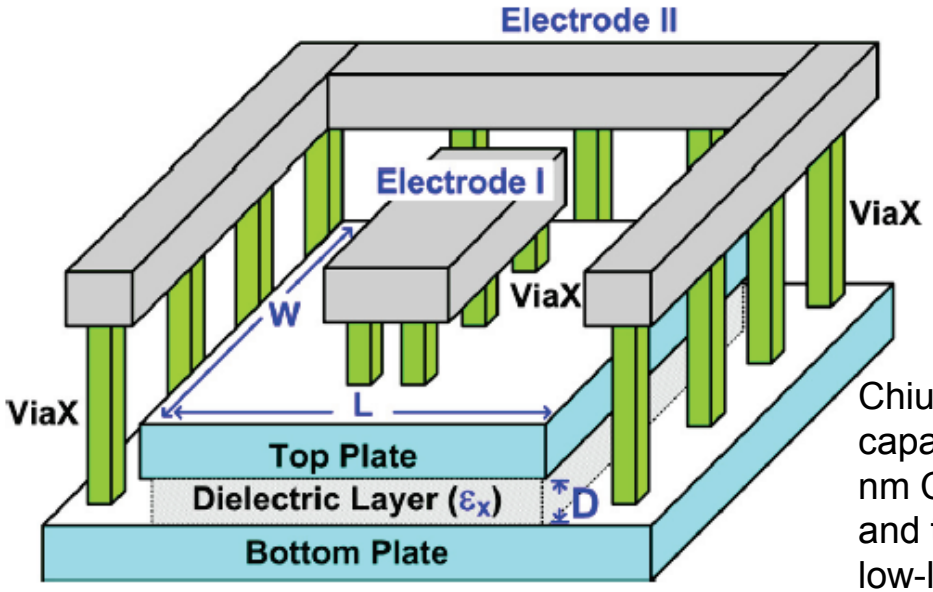
JÜLICH Forschungszentrum

QUANTENCOMPUTING

High level system model – Area estimation (65nm CMOS TSMC process, no optimization)



- Technology parameters from 65nm TSMC CMOS
  - Eff. cap density C_{den} = 1.75 fF/μm²
  - Transistor switch area A_{SW} = 0.37 μm²

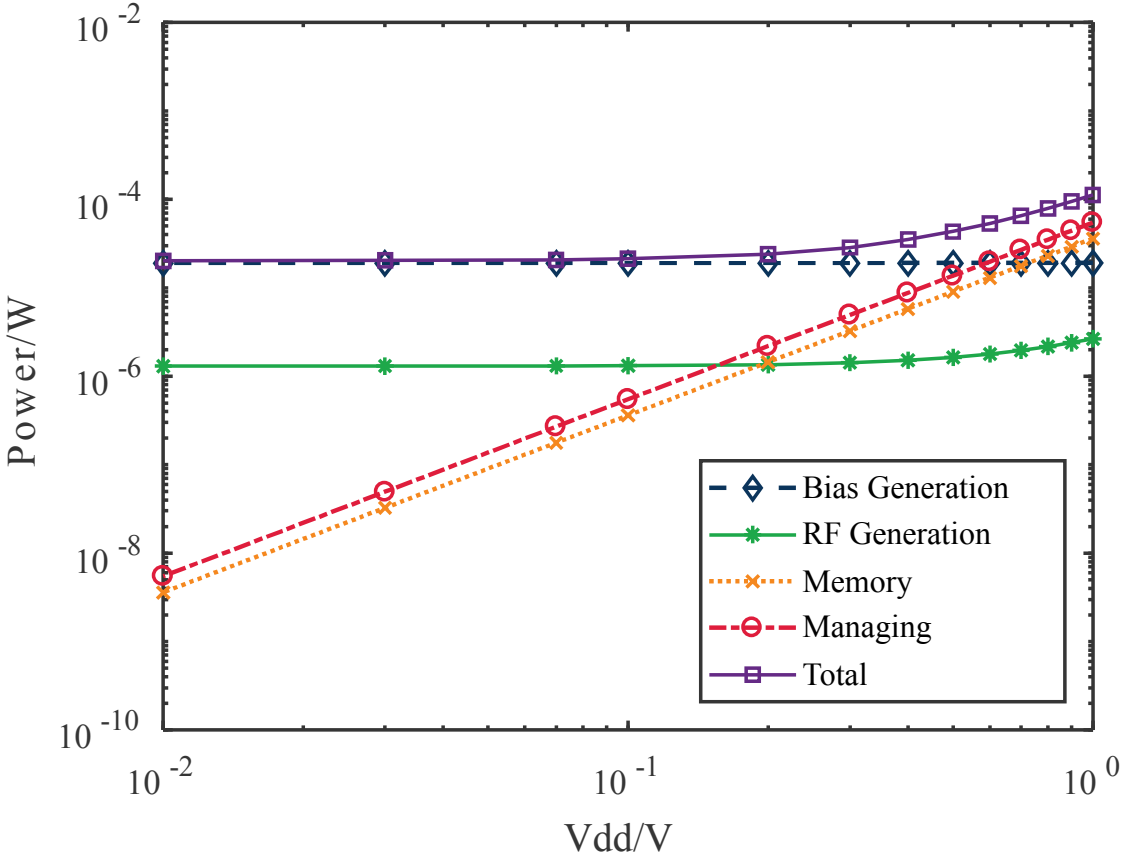


Chiu, Ker, „Metal-layer capacitors in the 65 nm CMOS process and the application for low-leakage power-rail ESD clamp circuit"

JÜLICH Forschungszentrum

QUANTENCOMPUTING

High level system model – Power estimation (65nm CMOS TSMC process, no optimization)



- Power estimation calculation base
  - DAC
    - $P_{DAC} = 0.5 \cdot f_{samp} \cdot V_{ref}^2 \cdot C_{tot}$
  - Logic
    - $P_{dig} = \sigma \cdot f_{samp} \cdot V_{dd}^2 \cdot C_{gate} + \sigma \cdot f_{samp} \cdot V_{dd}^2 \cdot Q_{sc} + V_{dd} \cdot I_{off}$
- Conservative simulation on activity rate σ :
- Memory: $\sigma_{mem} = 1.5 \cdot 10^{-4}$
- Logic: $\sigma_{logic} = 0.5$

QUANTENCOMPUTING

Bottom-up implementation – Bias voltage DAC (DC DAC) in charge-redistribution topology

- No static power dissipation
- Low thermal noise: $\bar{V}_N^2 = \frac{K_B \cdot T}{C}$
- Multiple output channel per DAC
- Loaded voltage divider:
  - Iterative charging to compensate voltage drop (leakage)
  - no output buffer needed
- Coarse setting reference voltage
  - reduce power and bits in charge redistribution part