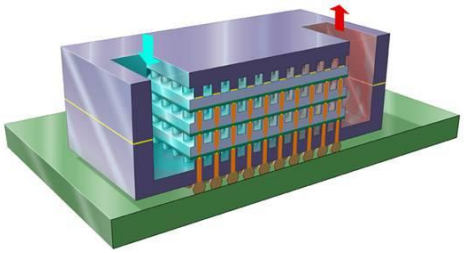
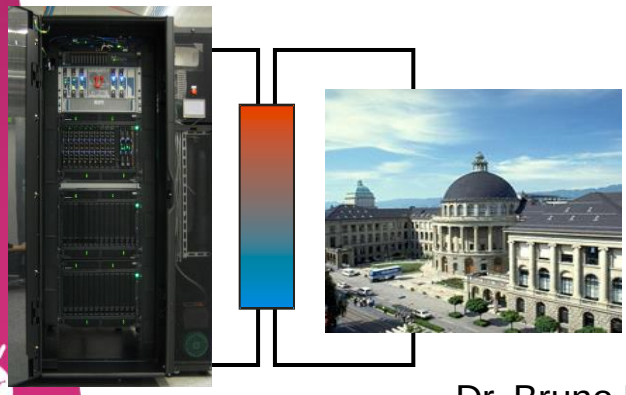
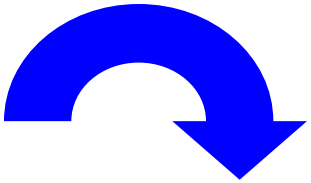
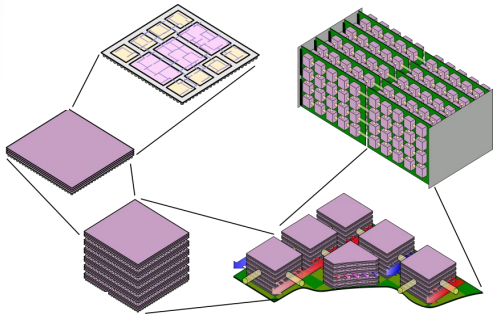
Roadmap towards Ultimately – Efficient Datacenters









Dr. Bruno Michel, Manager Advanced Thermal Packaging,
IBM Research Laboratory, Rüschlikon

Agenda

Paradigm Change 1:
From Cold Air Cooling to Hot Water Energy Re-Use

- Green Datacenter Drivers and Energy Trends
- Aquasar Zero Emission Datacenters: History and Vision
- SuperMUC Scaleup to 3PFLOPs
- From Hardware Cost to Total Cost of Ownership

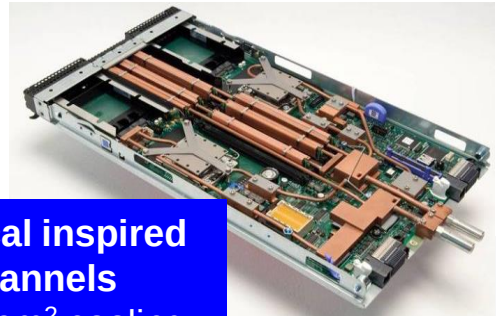
Paradigm Change 2:
From Performance to Efficiency

- From Maximal Performance per Chip to Performance per Joule
- Focus on Energy and Exergy
- Efficiency of Computer vs. Efficiency of Biological Brains

Paradigm Change 3:
From Areal Device Size Scaling to Volumetric Density Scaling

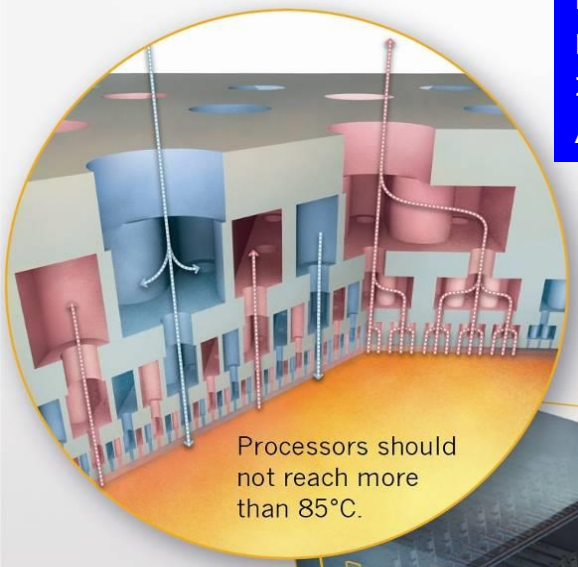
- The “Missing” Link between Density and Efficiency
- Interlayer Cooling and Electrochemical Chip Power Supply
- Link between Allometric Scaling and Rent’s Rule
- Towards Five-Dimensional Scaling

Paradigm Change 1: Hot-Water-Cooled Zero Emission Datacenters



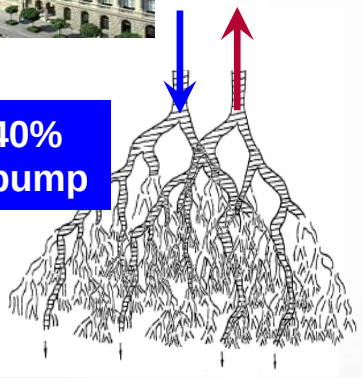
Aquasar Blade Center

Biological inspired microchannels
>700 W/cm^2 cooling
Allow <15ºC gradient



Energy Exceeds Hardware Cost; Keeps Rising

System at ETH eliminates 40% energy of datacenter heat pump



Water Pump

CMOS 80ºC

Collect Heat in Datacenter

Heat Exchanger

>3 years operation no failing parts

Direct Heat Use for Heating, Hot Water, Cooling, and Desalination

Water Out 65ºC

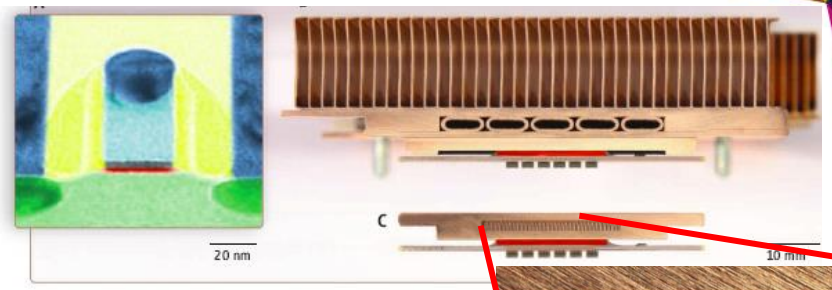
Eliminate ITC's 2% ww carbon emission; more than of airline industry

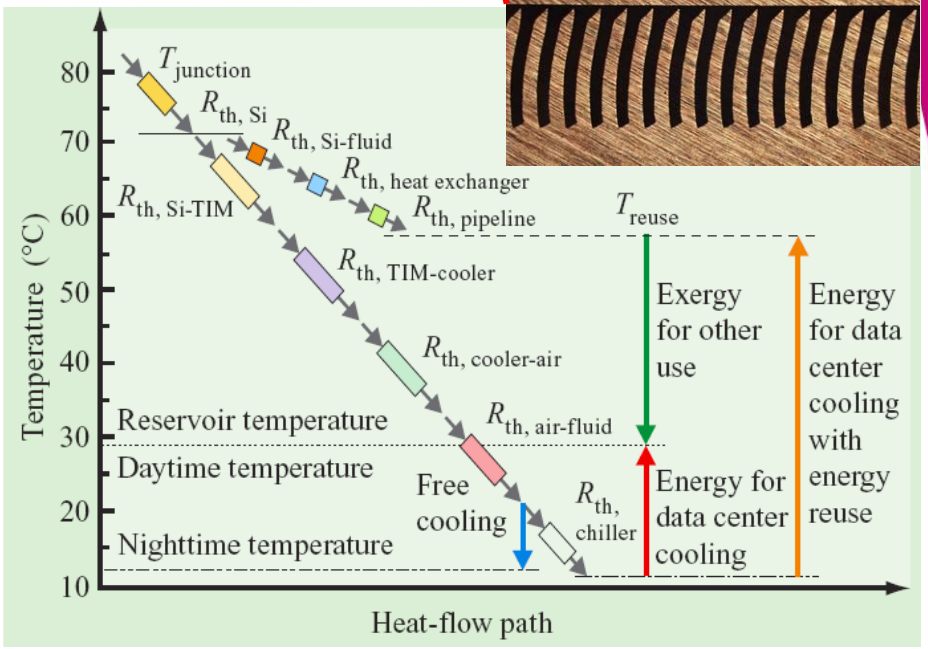
Water In 60ºC

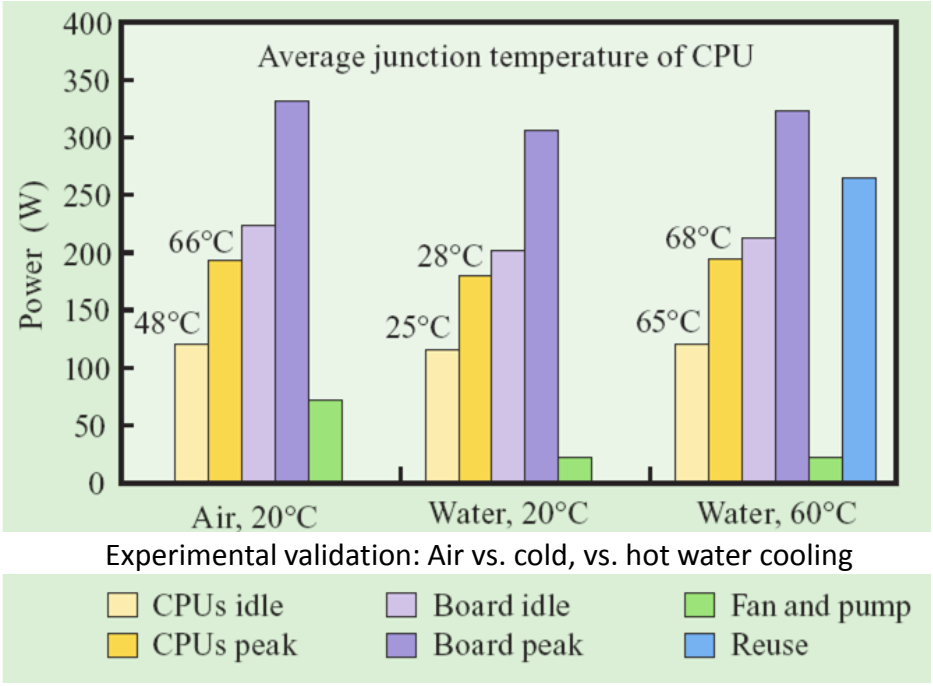
Economic value of heat reduces datacenter Total Cost of Ownership by 50-70%

Zero-Emission Data Centers

- **High-performance chip-level cooling** improves **energy efficiency** AND **reduces carbon emission:**
- Zero-emission **valuable in all climates**
  - Carbon footprint reduction
  - Cold and moderate climates: **energy savings** and **energy re-use**
  - Hot climates: Adsorpiton cooling, desalination
- **Facility Efficiency 7.9 TFLOP/gCO$_2$**





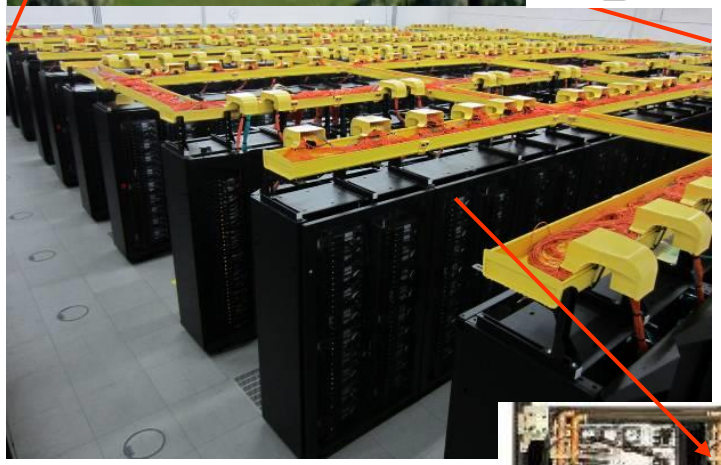


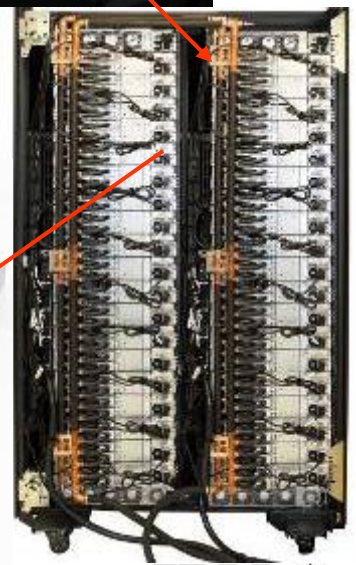
Experimental validation: Air vs. cold, vs. hot water cooling

- Europe: 5000 district heating systems
  - Distribute 6% of total thermal demand
  - Thermal energy from datacenters absorbed
- 3x smaller datacenter energy cost
- >3yrs operation, **no failing parts**

SuperMUC I and II

- Hot Water Cooled iDataPlex cluster with 3.2 / 2.9 PFlop/s peak / Rmax performance
  - ~20'000 CPUs / 160'000 Cores
- Energy Efficient **AND** Direct Heat Reuse
  - 4 MW Power, PUE 1.15, 90% heat for reuse
  - 40% less energy consumption
  - **Largest Computer in Europe (May 2012)**
  - **#1 in reuse list (ERE pending)**
- SuperMUC phase II announced
  - 3PF New more efficient compute hardware
  - Total machine power 7 MW (Phase I + II)
- System is part of the Partnership for Advanced Computing in Europe (PRACE) HPC infrastructure for researchers and industrial institutions throughout Europe
- SuperMUC is based on Aquasar Hot Water Cooling technology
- Largest universal High Performance CPU system

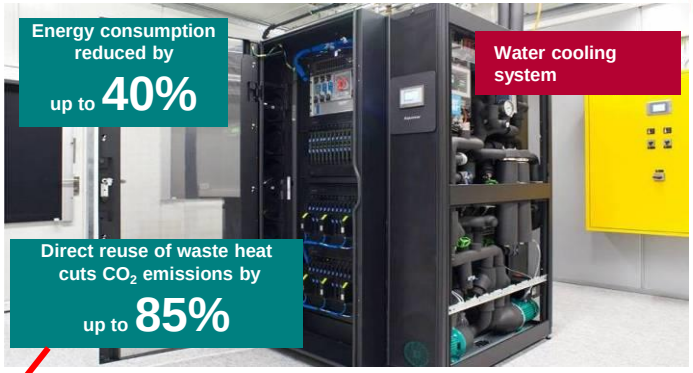


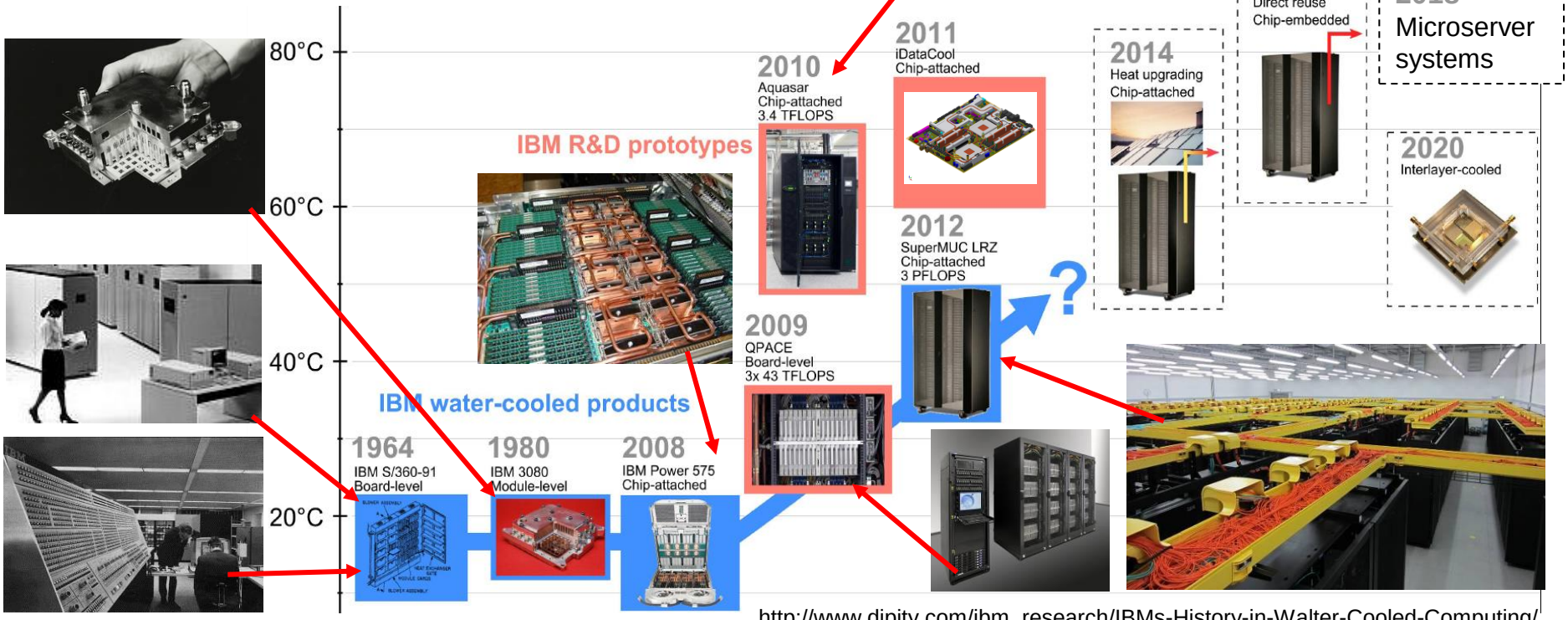


iDataPlex dx360 M4 board and rack

Aquasar / SuperMUC History and Vision

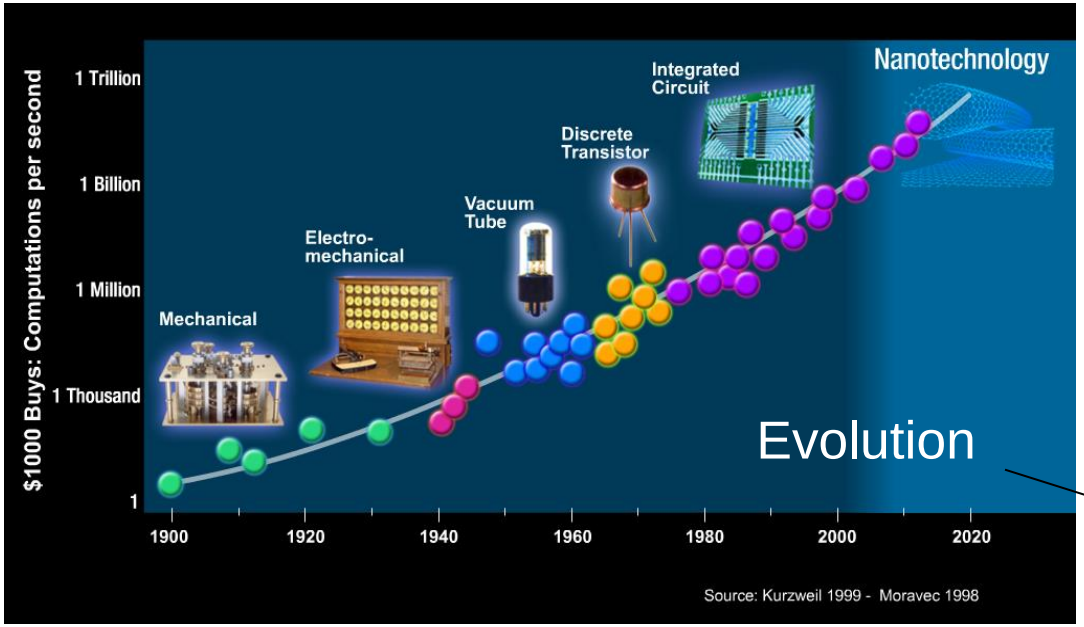
- Cold water module-level cooling for s3080 and p575
- QPACE prototype warm water cooling
- Aquasar with 65ºC chip attached hot water cooling
- iDataCool Prototype with 65ºC hot water cooling
- SuperMUC with 45ºC warm water cooling
- Embedded coolers allow hot water cooling (>65ºC)
- Prepare for interlayer cooled chip stacks (2020)





http://www.dipity.com/ibm_research/IBMs-History-in-Walter-Cooled-Computing/

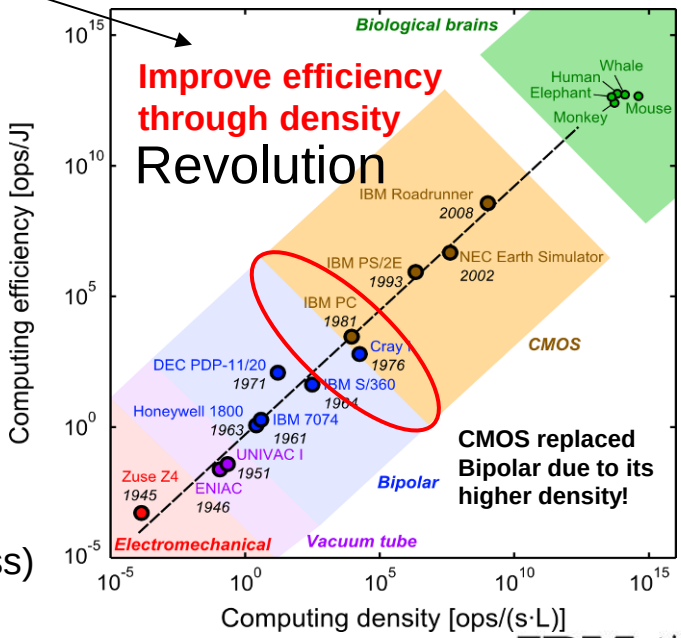
Paradigm Change 2 and 3: Revolution of Information Technology



- **Device centric viewpoint (left)**
  - ➔ Device performance dominates
    - Power depends on device performance
    - Evolution depends on better devices

vs.

- **Density centric viewpoint (below)**
  - ➔ Communication efficiency dominates
    - Power and memory proximity depend on size
    - Evolution depends on denser system
    - Dominant for large systems (>Peta-scale)



Information technology has prospered by making "bits" smaller.
➔ Smaller = *faster & cheaper (and more efficient)*

- Density and efficiency on log-log line
  - Brain is 10^4 times denser AND 10^4 times more efficient
- Independent of switch technology
  - No jumps mechanical – tube – bipolar – CMOS – neuron
- Communication as main bottleneck
  - Memory proximity lost in current computers (1300 clock access)
  - Detrimental for efficiency

Why Size Matters so much for Computers

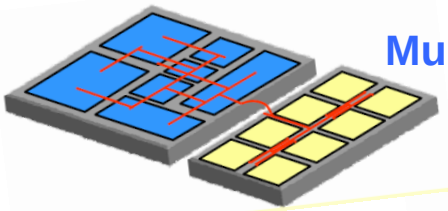
- Today's systems: Transistors occupy only 1 ppm of the system volume ~1'000'000ppm power supply & cooling

➔ Never before devices occupied a smaller volume fraction

- PC AT used about same amount for computation and communication
  - Since then processor became 10'000 times better
  - PCB and C4 interface only improved 100 times
- Majority of Energy used for data transport in current computers
  - 99% communication and 1% computation
  - 1300 clock cycles needed for main memory access
- Major reason C4 bottleneck that creates "memory wall"
  - 3D integration moves main memory into chip stack
  - "Cooling wall" is solved by interlayer cooled chip stacks
- Brain serves as example for dense and efficient computing
  - 3D integration and memory proximity is key for efficiency
  - Brain has similar Rentian slope as microprocessors
  - Brain communication density lower for 1 neuron = 1000 transistors

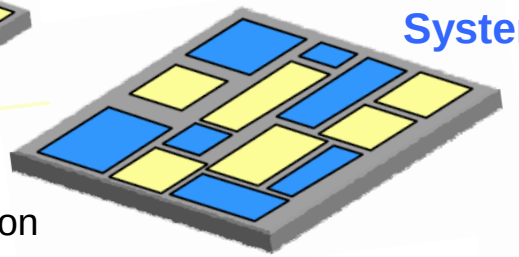
Paradigm Change: Vertical Integration

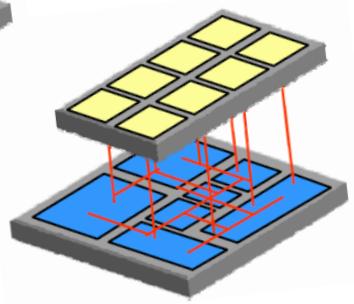
Brain: synapse network



Multi-Chip Design

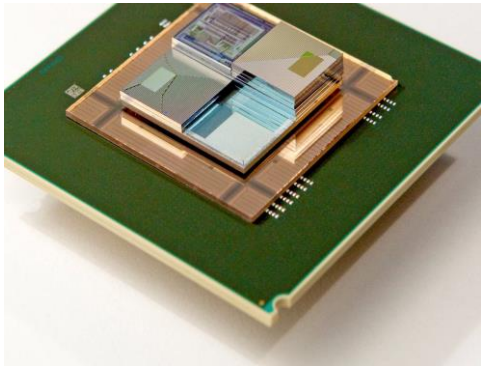
System on Chip

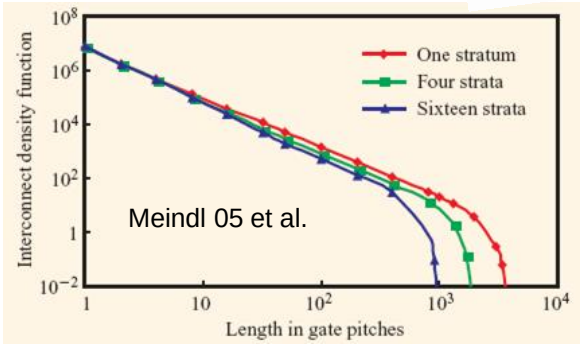




Global wire length reduction

3D Integration



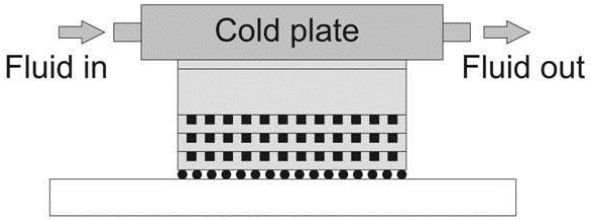


Benefits:

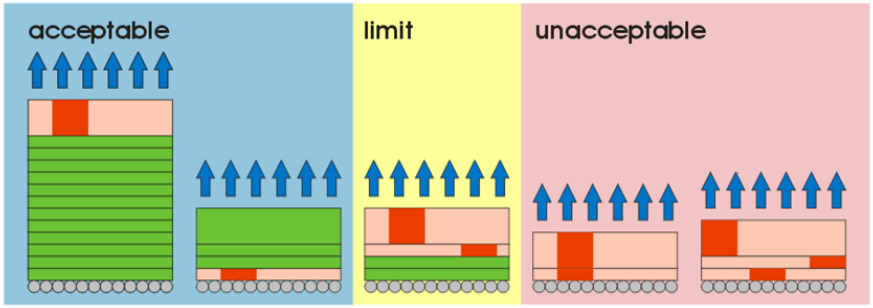
- High core-cache bandwidth
- Separation of technologies
- Reduction in wire length

→ Equivalent to two generations of scaling

→ No impact on software development



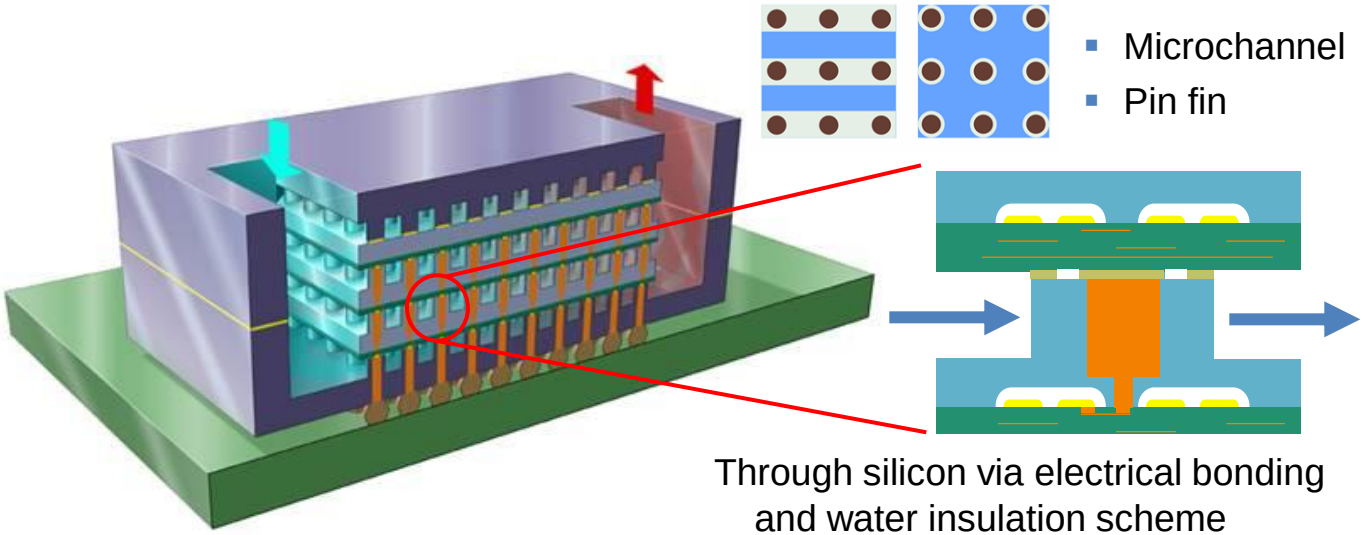
Microchannel back-side heat removal



BUT: Heat removal limit constrains electrical design

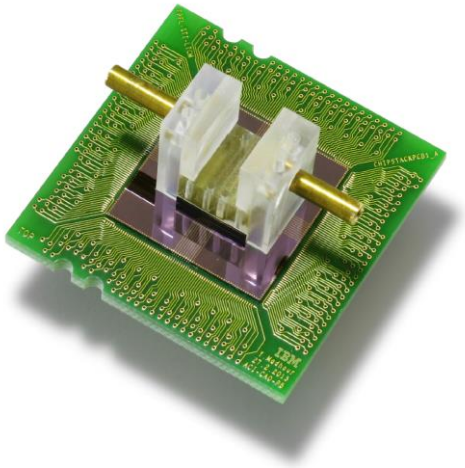
Scalable Heat Removal by Interlayer Cooling

- 3D integration requires interlayer cooling for stacked logic chips
- Bonding scheme to isolate electrical interconnects from coolant

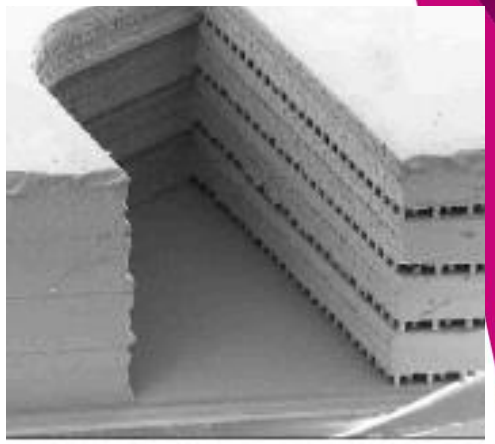


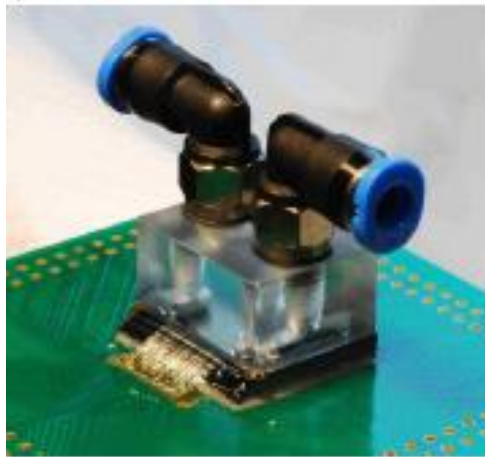
Through silicon via electrical bonding and water insulation scheme

- A large fraction of energy in computers is spent for data transport
- Shrinking computers saves energy



cross-section through fluid port and cavities

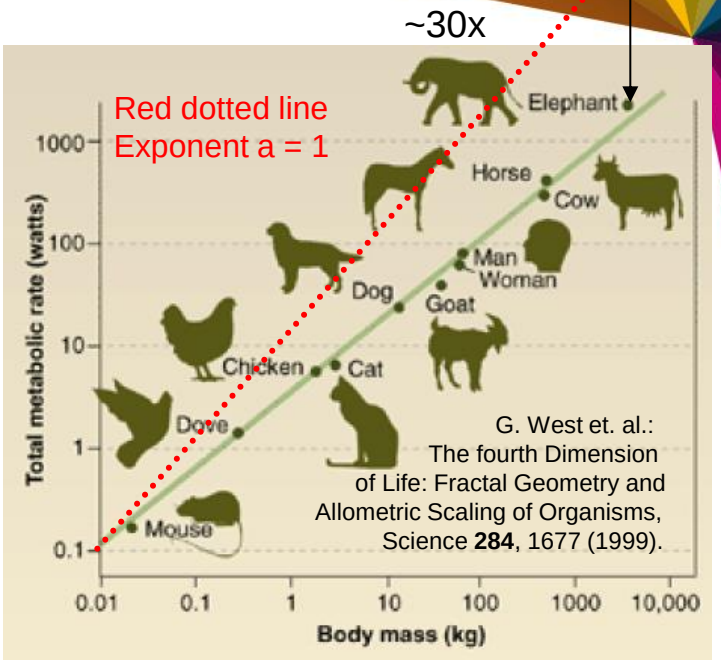


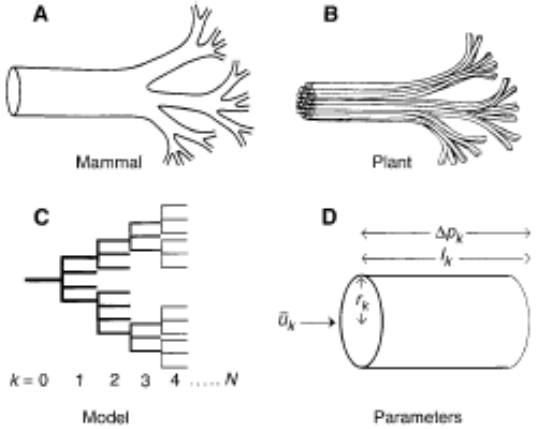


Test vehicle with fluid manifold and connection

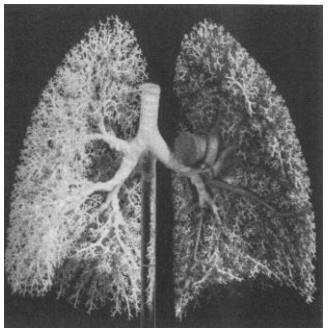
Allometric Scaling in Biology

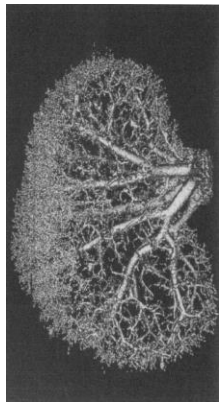
- Increasing size in biology: Scaling behavior: $R \propto M^a$
- Metabolic rate with increasing organism size
- Kleiber (1932): Scaling exponent is $R \propto M^{3/4}$
- West (1997): Exponent ¾ explained by hierarchically branched supply network
- Kleiber, Physiological Reviews 27 (1947) 511
- Schmidt, Why is Animal Size important? (1984)
- West et al., Science 276 (1997), 122
- Mackenzie, Science 284 (1999), 1607





- From Engineering to Nature ➔ and to Engineering
  - Lung (blood in+out, air in/out)
  - Kidney (blood in+out, water out)
  - Tree (roots, leaves)
  - River basins
  - Drying mud
- Human built
  - Dwellings and Cities
  - ~~Computers~~



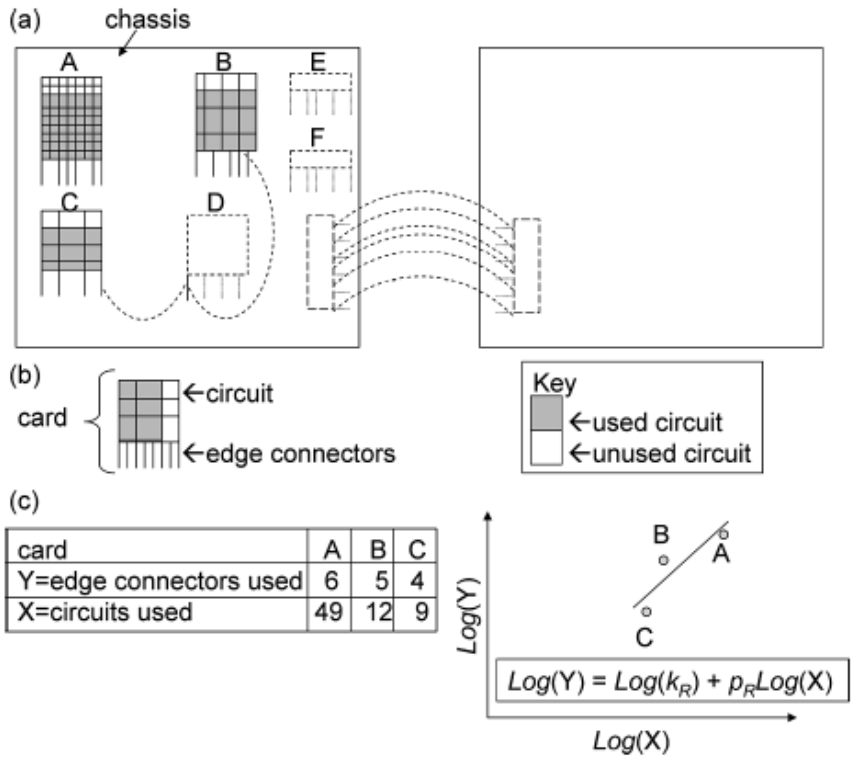


Biology builds dense and efficient complex systems
Branched networks as important as genetic code!

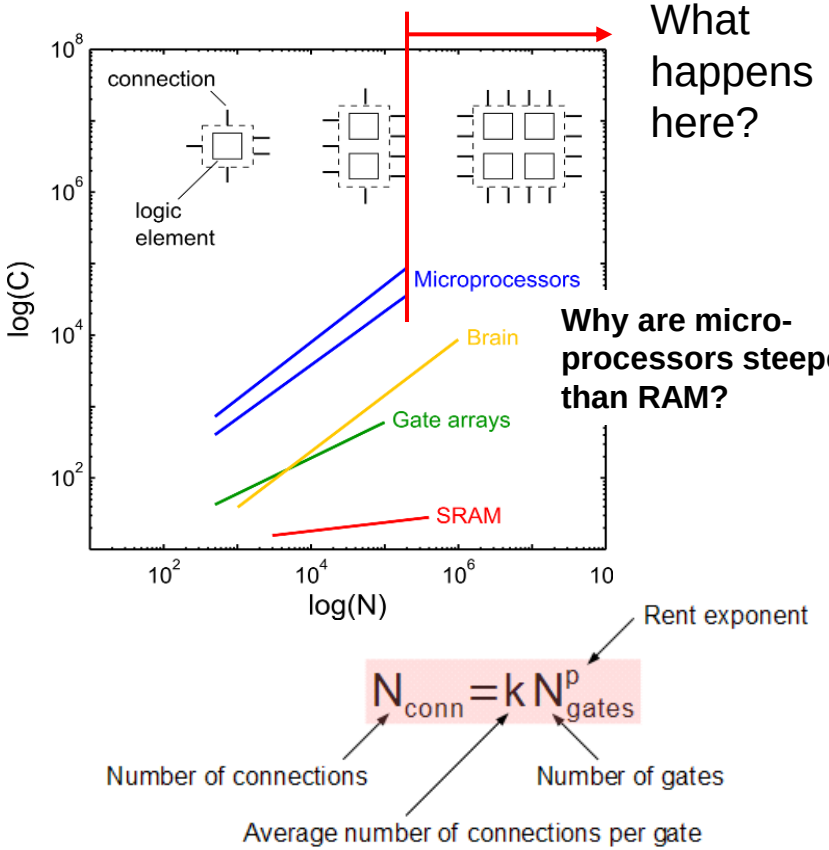
- Bejan, "From Engineering to Nature", Cambridge University Press, 2000. ➔ Constructal Design – Ubiquitous in Nature

Describing Communication Needs with Rent's Rule

- In 1960, E.F. Rent of IBM counted the number of socket pins (edge connectors) per card required to communicate with all circuits (logic blocks) on that card.

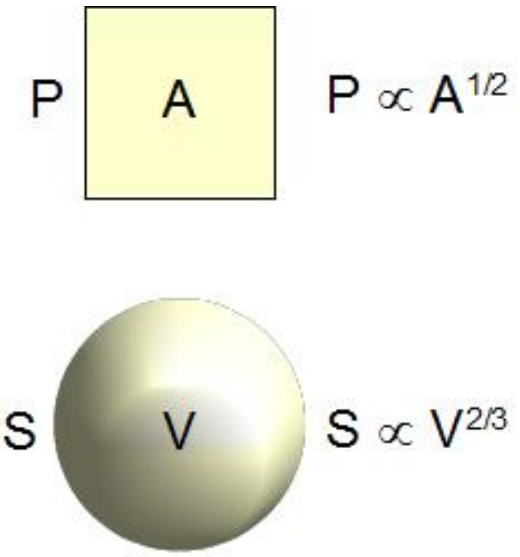


card	A	B	C
Y=edge connectors used	6	5	4
X=circuits used	49	12	9



Of Mice and Men…
and Elephants and Whales

- Biological (allometric) volumetric scaling
  - Allometric scaling: Exponent 0.75 ➔ 4 D scaling
  - Why? Chemical power supply and hierarchical supply networks
  - Fluid pressure drop scales 4-dimensional
- Linear vs. hierarchical blood supply
- Virtual 4th dimension
  - Human volume 0.064 m^3 with skin surface 1 m^2
  - Internal surface (lung, intestine, kidney) > 600 m^2
  - Human uses a virtual volume of 1000 m^3



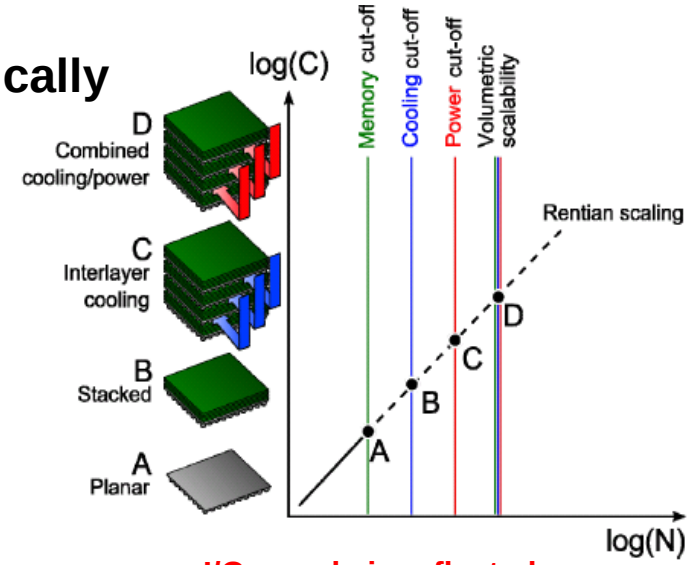
Scaling $\propto N^{(D-1)/D}$ corresponds to scaling in the D^{th} dimension!

General
$S \propto N^{(D-1)/D}$ for D^{th} dimensional geometry

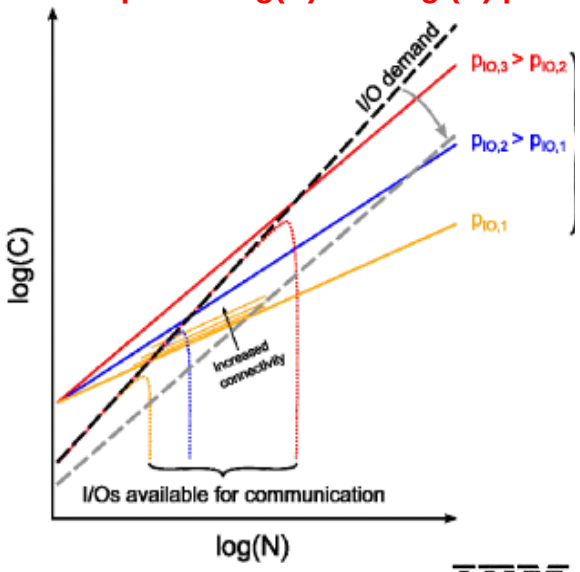
• Mouse:	body weight 10 g	food per day 10 g	
• Elephant:	body weight 10 t	food per day 300 kg	
•	haul a 1 t log over 1 km	haul a 1 g straw over 1 km	
•	(≈ 1'000'000 Mice)	(≈ 30'000 mice)	➔ 31 times lower metabolic rate

Density Improves Efficiency

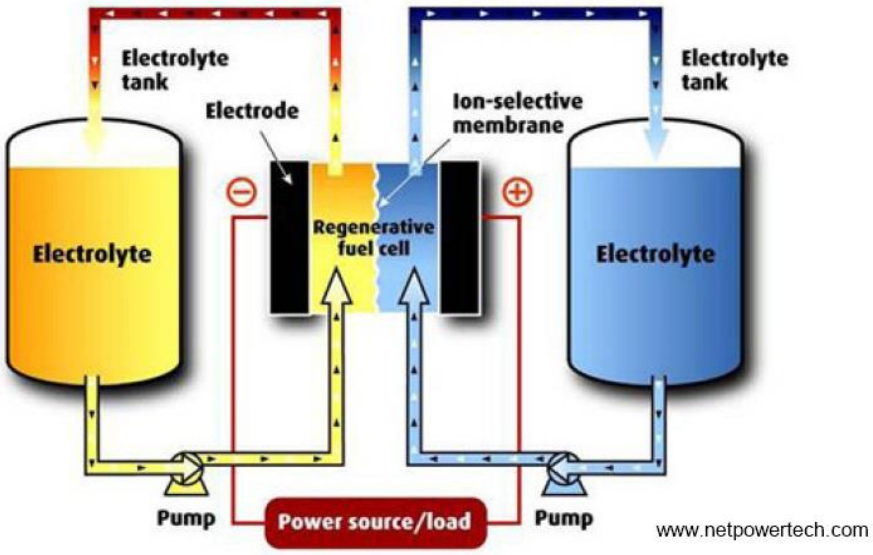
- **Communication energy dominates quadratically**
  - Power and memory proximity dependent on wire length
  - Communication energy scales faster than size
- **Memory proximity restored in chip stack**
  - Main memory in stack – no cache necessary
  - Interlayer cooling removes cooling wall
  - Electrochemical power supply removes power wall
- **Reach density AND efficiency of brain**
  - CMOS technology can reach sufficient density
- **Key volumetric scaling laws**
  - Device count AND power demand scale with volume
  - Communication AND power supply scale with surface
  - Large-system performance scales with Hypersurface / Hypervolume = 1-D / D
- **Biological (allometric) volumetric scaling**
  - Allometric scaling: Exponent 0.75 ➔ 4 D scaling
  - Why? Chemical power supply and hierarchical supply networks
  - Fluid pressure drop scales 4-dimensional



I/O supply is reflected as Slope on Log(C) vs. Log (N) plot

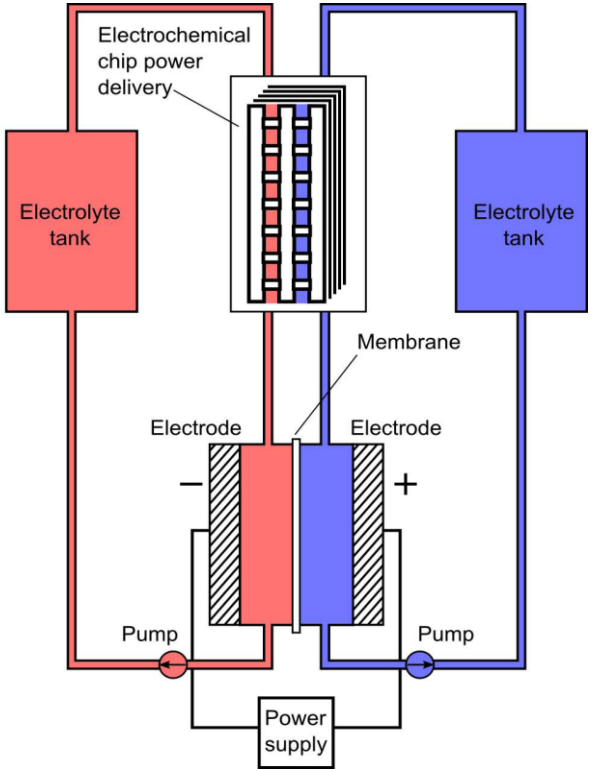


Electrochemical Redox Flow Batteries



- **Characteristics**
- Soluble redox species
- Inert electrodes
- Independent energy and power properties
- Single charge and discharge unit

- **Technology benefits**
- No changes in electrode active surface area
- Deep discharge and high power possible
- No electrode lifetime limitations

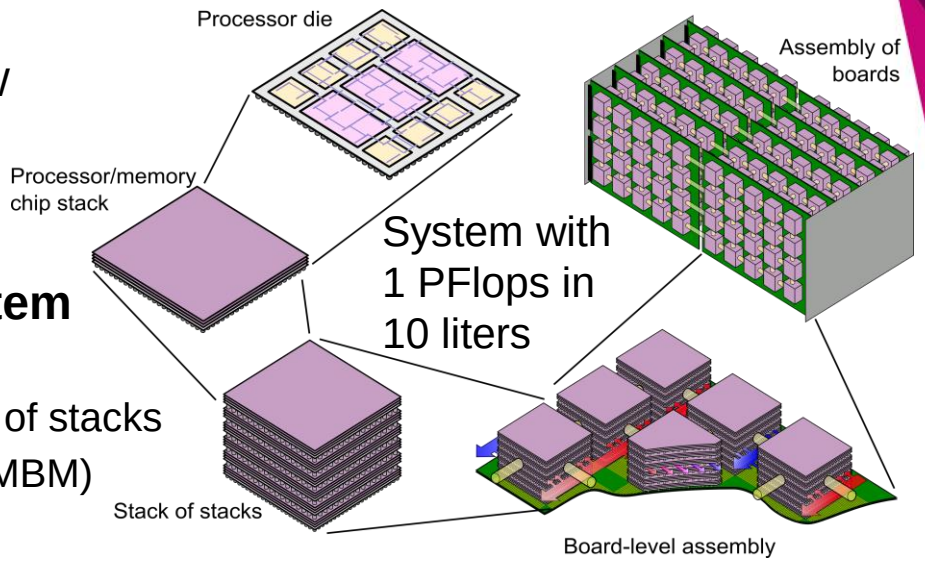


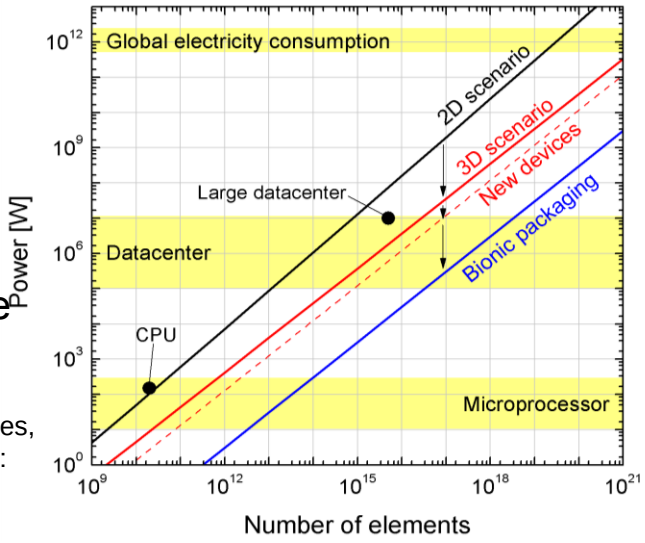
Electrochemical chip power supply

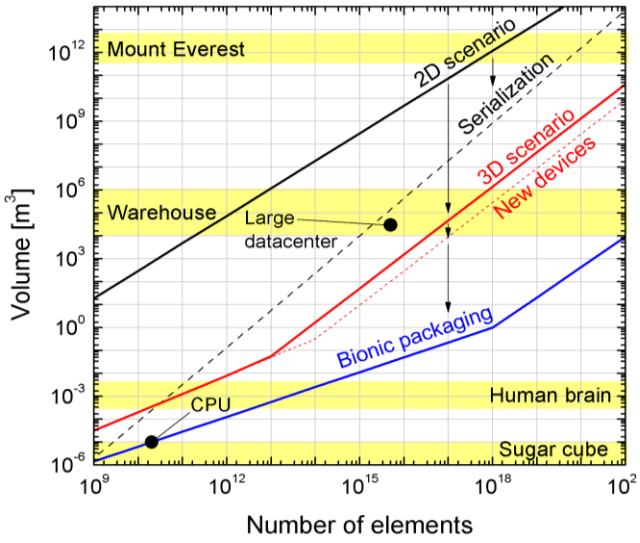
- Single macroscopic charging unit
- Multiple chip-level discharge units
- Satisfies congruent demand for power delivery and heat removal

Scaling to 1 PFlops in 10 Liters

- **Efficiency comparison**
  - 1PFlops system currently consumes ~10MW
  - 0.1 PF ultra-dense system consumes 5kW
  - Conventional power supply scales causes power supply wall
- **Extreme 3D 1PFlops ultra-dense system**
  - Stack ~10 layers of memory on logic
  - Stack several memory-logic stacks to stack of stacks
  - Combine several blocks of stacks to MCM (MBM)
  - Combine MCMs to high density 3D system
- **Key enabling technologies**
  - Interlayer cooling
  - Electrochemical chip power supply
- **Impact**
  - 5'000x smaller power
  - 50'000'000x smaller volume
  - Scalability until zetascale



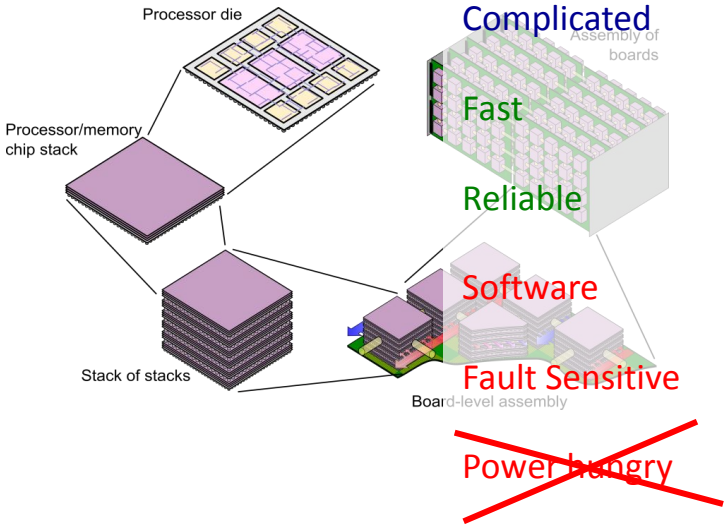




P. Ruch, T. Brunschwiler, W. Escher, S. Paredes, and B. Michel, "Towards 5 dimensional scaling: How density improves efficiency in future computers", IBM J. Res. Develop. 55 (5, Centennial Issue), 15:1-15:13 (2011).

Outlook Brain Inspired Computing

- Stepwise introduction of brain inspired concepts: Form – Function – Material
- Step 1 (Form): Brain inspired packaging with classical CMOS ➔ Now
- Step 2 (Function): Brain-inspired, non-von Neumann architecture ➔ Later
- Step 3 (Material): Artificial Neurons, or DNA computing … ➔ Far in the future
- Each step has to provide benefit when applied alone
- Bionic packaging equally supports von Neumann and non von Neumann architecture
- Models show a maximal efficiency gain of 5‘000 for radical 3D bionic packaging
- Relative importance of Steps 1 and 2 not clear



Computer

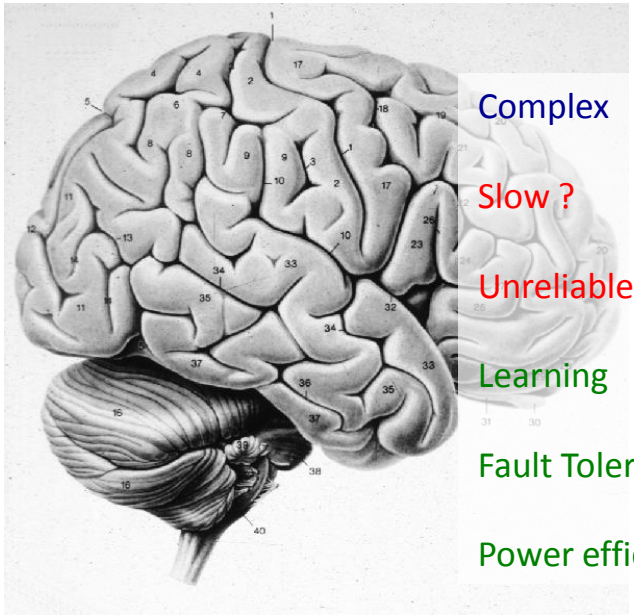
vs.

Human

in

Chess and

Jeopardy



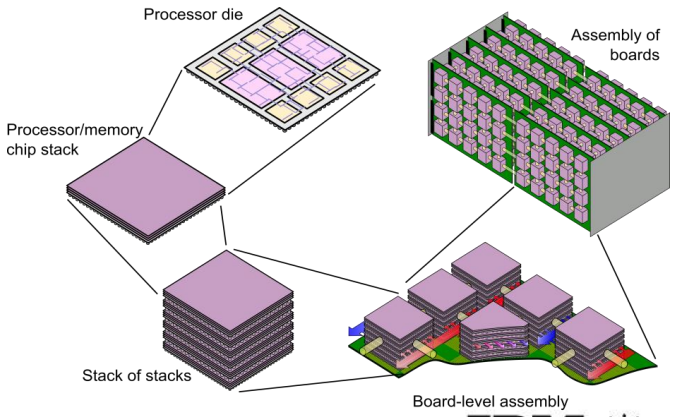
Summary

- **Paradigm changes reuse and efficiency**
  - Energy will cost more than servers
  - Liquid cooling and heat re-use: Aquasar / SuperMUC
  - Reduce >85%, save 40% energy and reduce energy cost by 2-3 x
  - Efficiency / carbon footprint and not performance is key

- **Moore's law goes 3D**
  - Stacking of layers allows extension of Moore's law
  - Interlayer cooled 3D stacks
  - Areal scaling is "almost dead" ➔ long live volumetric density scaling!

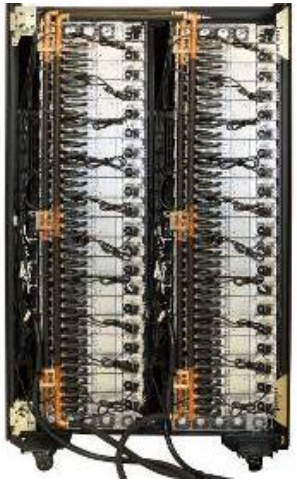
- **Volumetric Density Scaling and Bionic Packaging**
  - Functional density and connectivity of Human brain
  - Cooling + power delivery ➔ Bionic packaging
  - Shrink SuperMUC to 10liters: 5000x better efficiency
  - New scaling roadmap for next 15 years

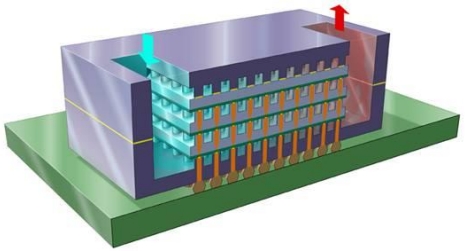
  - Next Steps: Microserver and REPCOOL Projects
  - Synergy with solar power ➔ Smart Planet

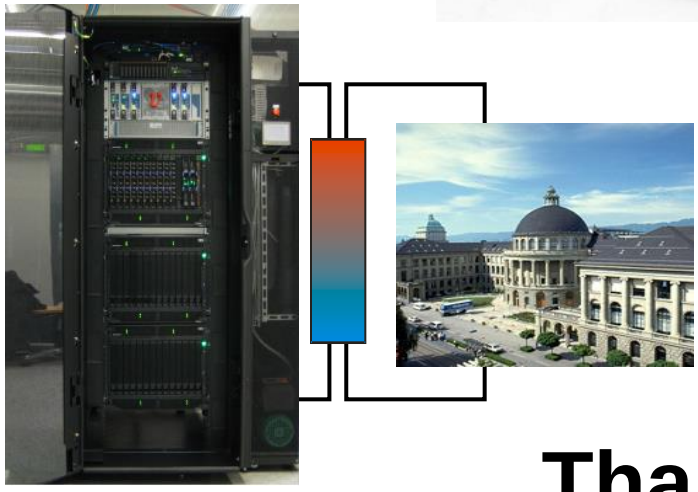


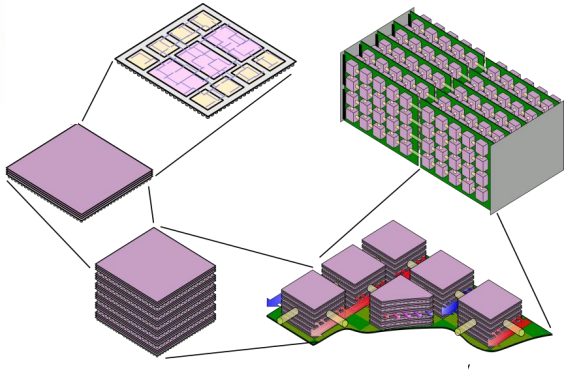
Acknowledgment:

- Ingmar Meijer, Patrick Ruch, Thomas Brunschwiler, Stephan Paredes, Werner Escher, Yassir Madhour, Jeff Ong, Gerd Schlottig.
- PSI Tobias Rupp and Thomas Schmidt
- ETH Severin Zimmermann, Adrian Renfer, Manish Tiwari, Dimos Poulikakos
- EPFL Yassir Madhour, John Thome, and Yussuf Leblebici
- Many more for Aquasar and SuperMUC design and build
- Funding: IBM FOAK Program, IBM Research, CCEM Aquasar project, Nanotera CMOSAIC project
- SNF Sinergia project REPCOOL









Thank you for Your attention

DOME – Research Phase for SKA (SKA = Square Kilometer Array)

The SKA will be the largest and most sensitive radio telescope ever built.

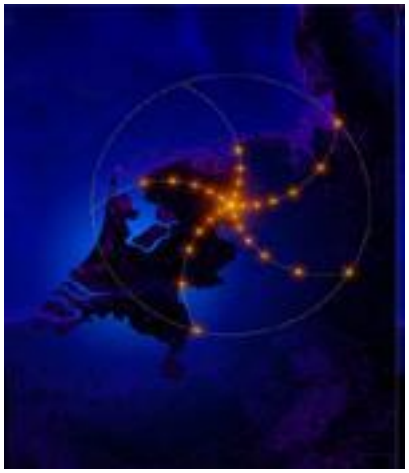
10’000’s of Antennas operational at 2020-24 with frequency ranges 70 MHz to 10 GHz will generate huge amounts of data, which needs to be *transported, analyzed, stored and retrieved*

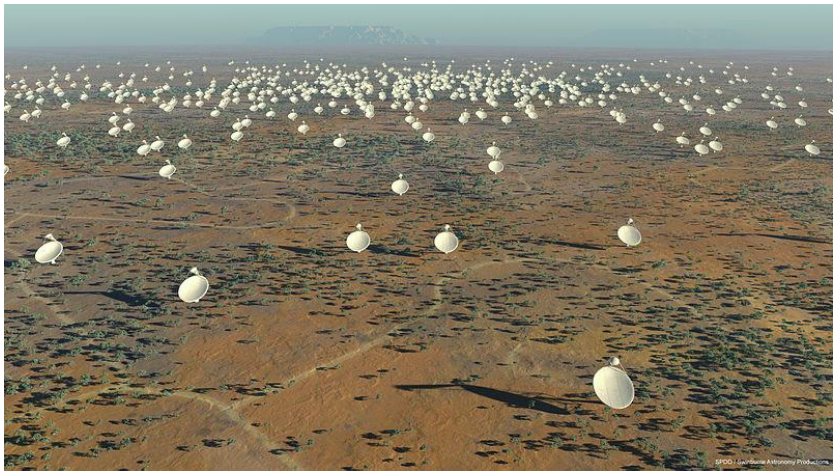
A true Exascale Analytics Challenge!

Every day, the antennas will gather 14 exabytes and store about one petabyte.

DOME is a research phase project before start of SKA deployment in 2017

- 5 year collaboration between ASTRON (NL) and IBM, started Feb 2012
- Co-funded by Dutch government with double digit Euro Millions (just for work, not for license)

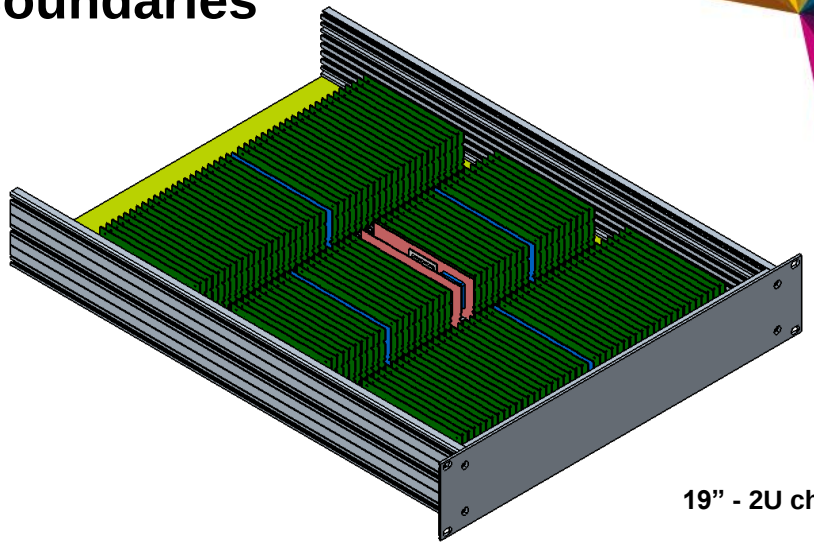




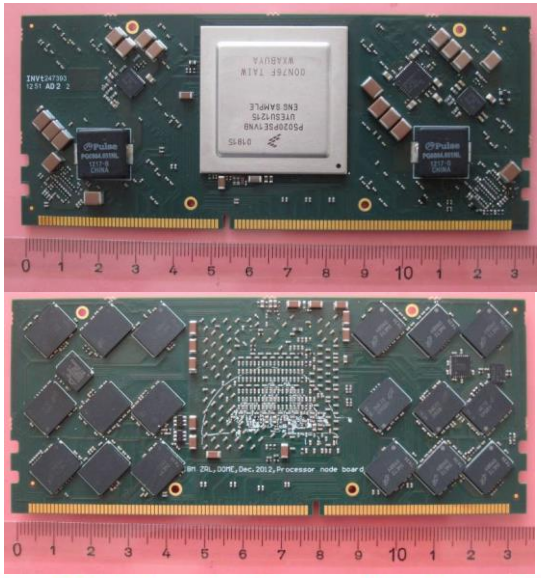
Microserver: System Definition and Boundaries

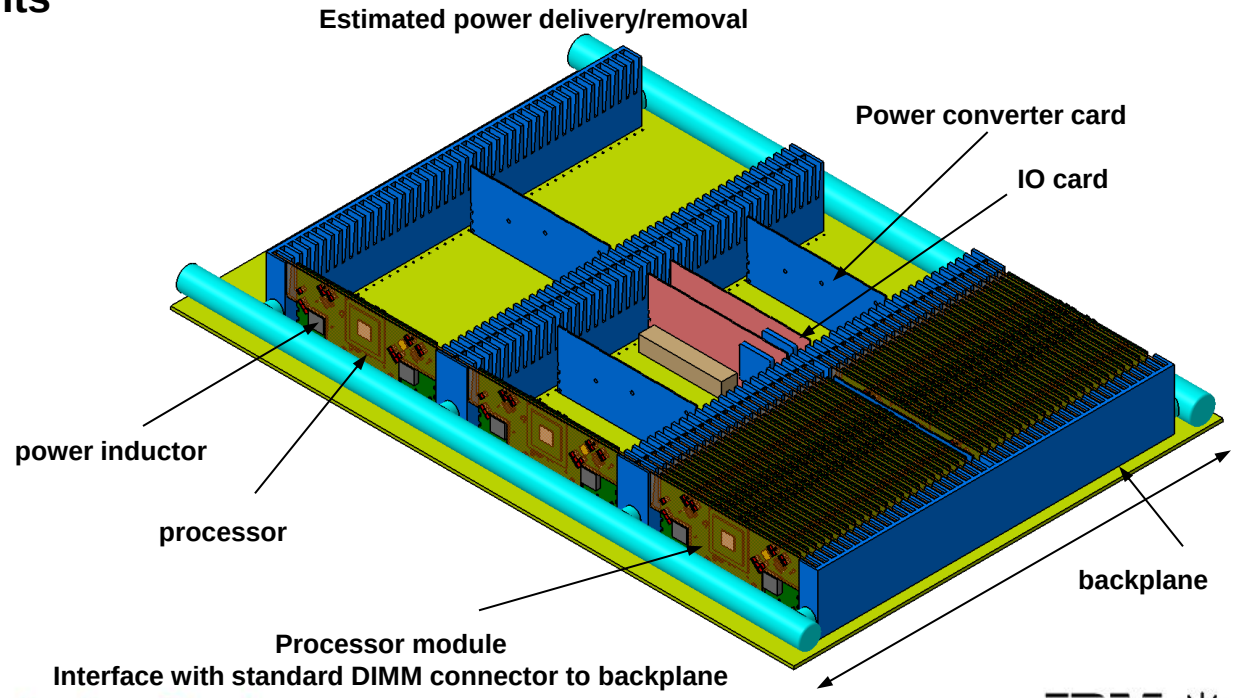
- High density µ-server system in 2U 19” drawer
- Processor modules in DIMM form factor
- Power converter modules
- IO modules
- Power supply (220 V AC to 12V DC)
- Backplane ~8 layer

➔**Minimal spacing between modules using standard components**



19” - 2U chassis



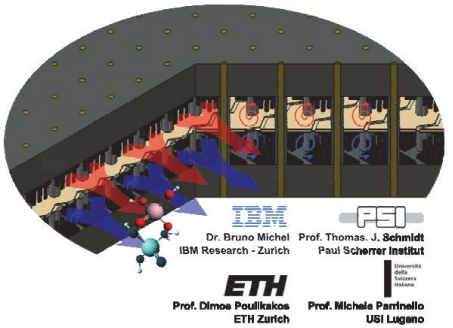


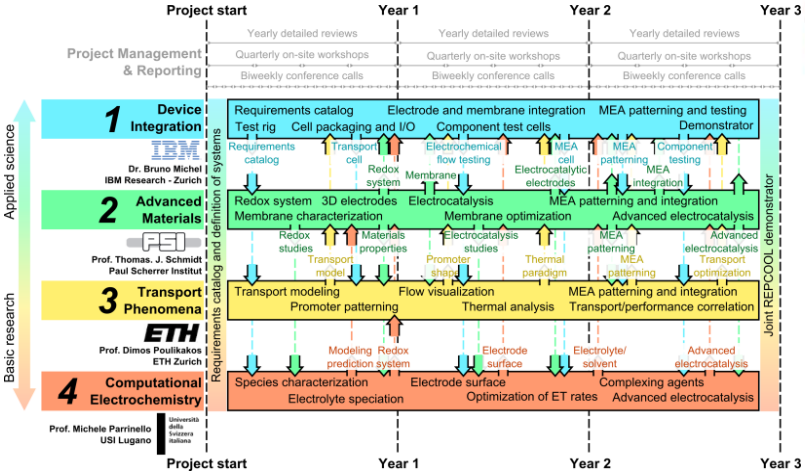
21

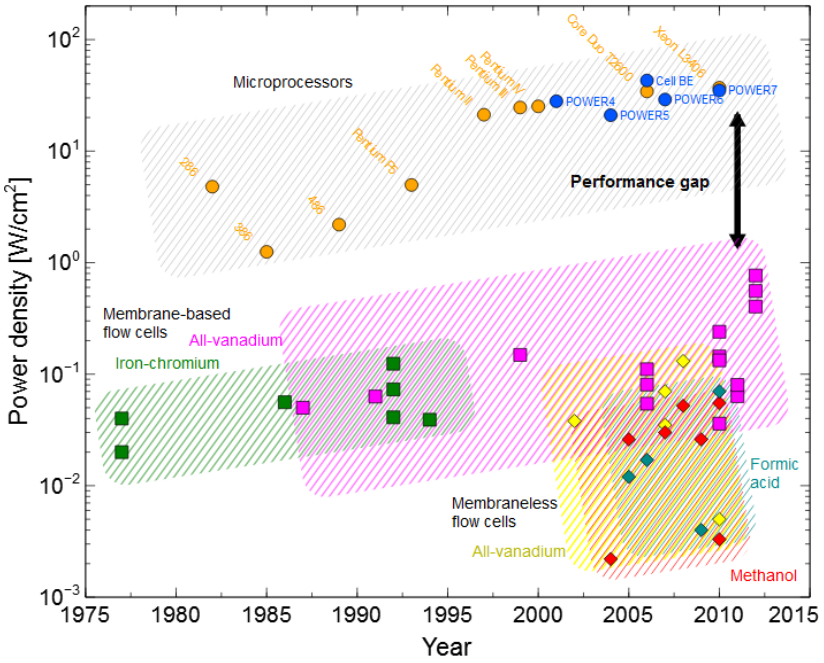
SNF Sinergia REPCOOL Project

- Redox flow electrochemistry for power delivery and cooling
- 3 year project funded by Swiss National Science Foundation
- 4 contributing sub-project
- Device Integration (IBM, Rüschlikon)
- Advanced Electrochemical Materials (PSI, Villigen)
- Microscale Transport Phenomena (ETH, Zürich)
- Computational Electrochemistry (USI, Lugano)

Main research target cover power density gap
for electrochemical power delivery in microchannels



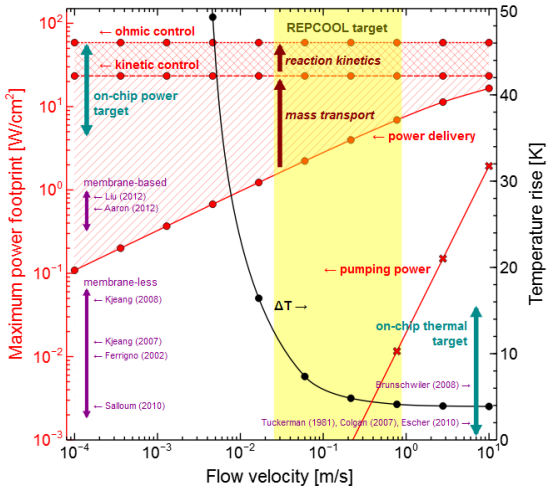




Approach:

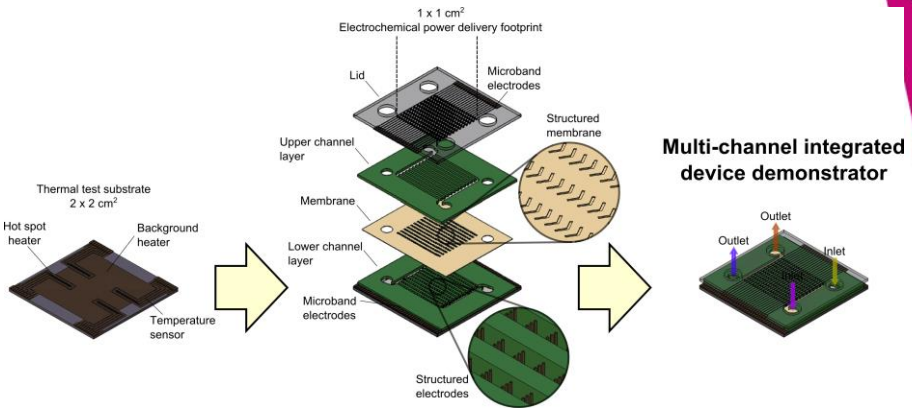
- Reaction Kinetics
- Mass Transport
- Electrochemical Katalysis
- Miniaturization

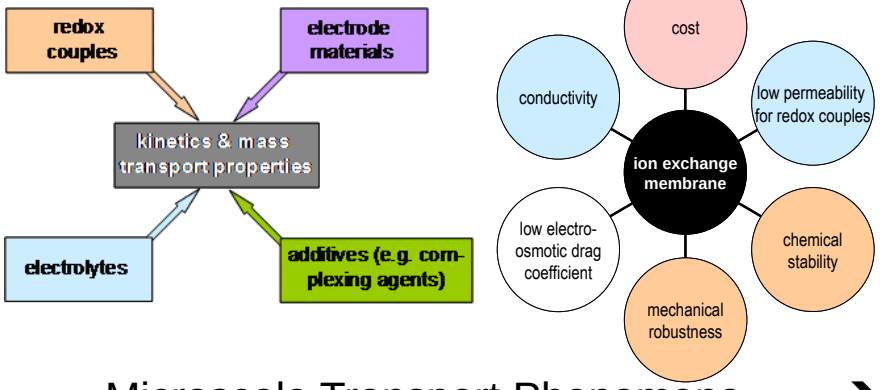
- Project Start Press Release in a few weeks



REPCOOL Sub-Projects

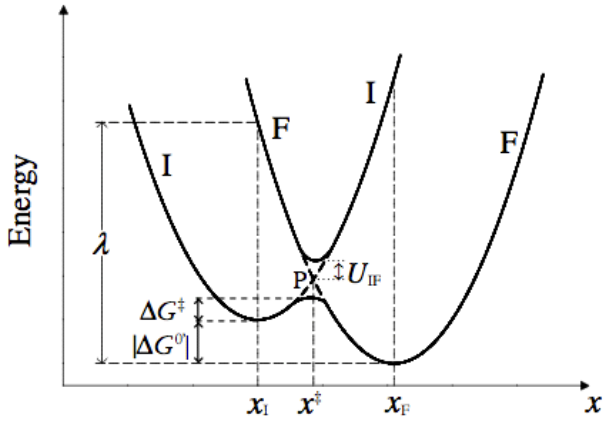
- Device Integration for Simultaneous Cooling and Power Delivery ➔

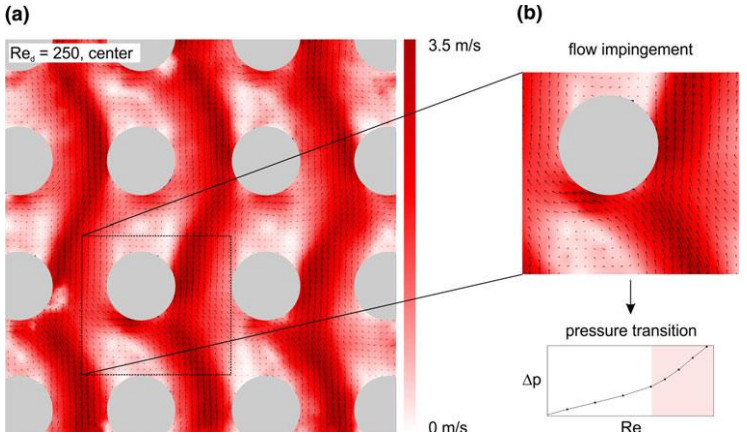




➔ Advanced Electrochemical Materials and Membranes

- Microscale Transport Phenomena ➔





➔ Computational Electro-Chemistry

Backup Slides

Sustainability and Economic Roadmap …
The Stairway to Heaven

Heaven

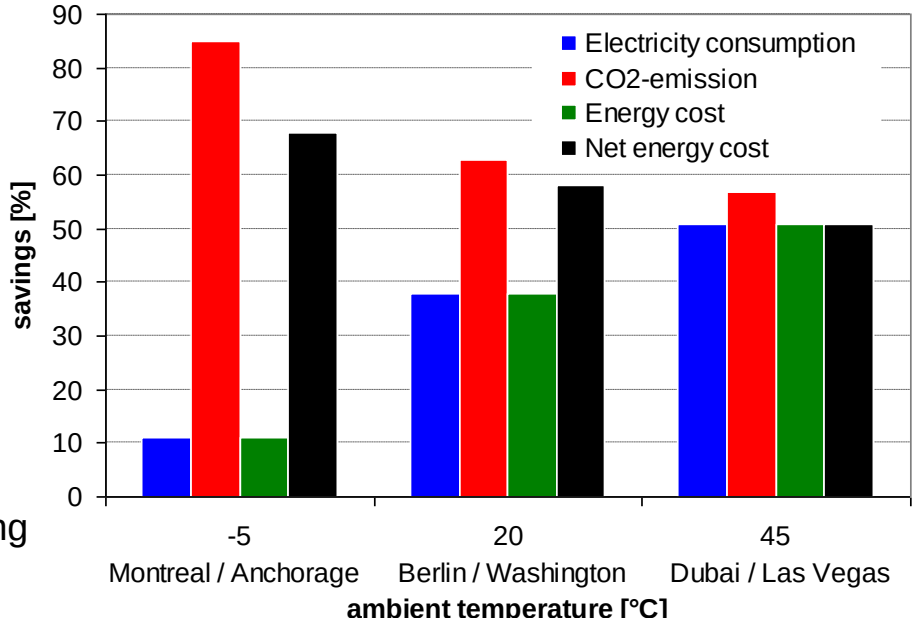
- Improved cooling helps to additionally **reduce the energy consumption of the computer** (address the quantity of 1 instead of the 0.1!)
- Improved cooling pays a part of the energy bill in all climates; requires higher temperatures which is not visible in ERE and ERF!
  - Inclusion of adsorption chillers
  - Heating of a swimming pool at 20ºC is easy driving an adsorption chiller at 70ºC is hard
- Improved cooling **pays a part of the energy bill** in cold climates
  - Goal to reach ERF = 1 at PUE = 1.1 (USV 0.05 and all pumps 0.05)
- Improved cooling **pays for itself** (PUE=1.1, ERE < 0.5, ERF > 0.5)
  - Aquasar PUE=1.15, ERE = 0.4, ERF 0.75
  - Invest into pumping (PUE 1.1 ➔ 1.15 to get ERF from 0 ➔ 0.75)
  - Selling of energy for reuse allows reasonable ROI on investment
- Improved cooling is (almost) **free** (PUE = 1.1, ERE=1.1, ERF = 0)
- We **pay** for cooling (50% overall cost = PUE 2)
  - Usually energy consumption of cooling tower fans is not included in PUE!

Hell

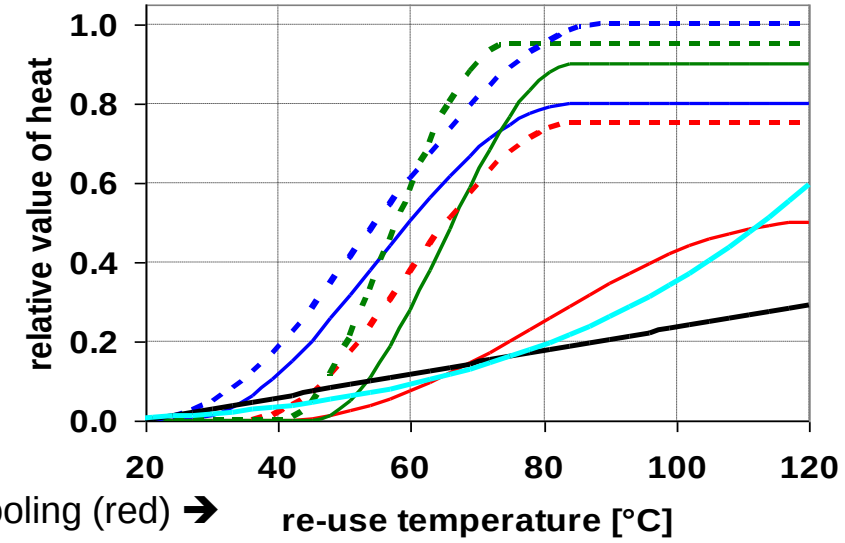
PUE = Total Energy / IT Energy = (Cooling + Power + Lighting + IT) / IT
ERE = (Cooling + Power + Lighting + IT – Reuse) / IT ERF = Reuse Energy / Total Energy

“Value” of Heat

- Basic concept
  - Cold and moderate climates: **energy savings** and **energy re-use**
  - Hot climates: free cooling during hottest day recorded on earth (57ºC)
- Longer term developments
  - Drive desalination with datacenter heat: Project with Egypt on use of solar heat
  - Drive adsorption chiller with hot water cooling
  - Cool air cooled components in hot climates

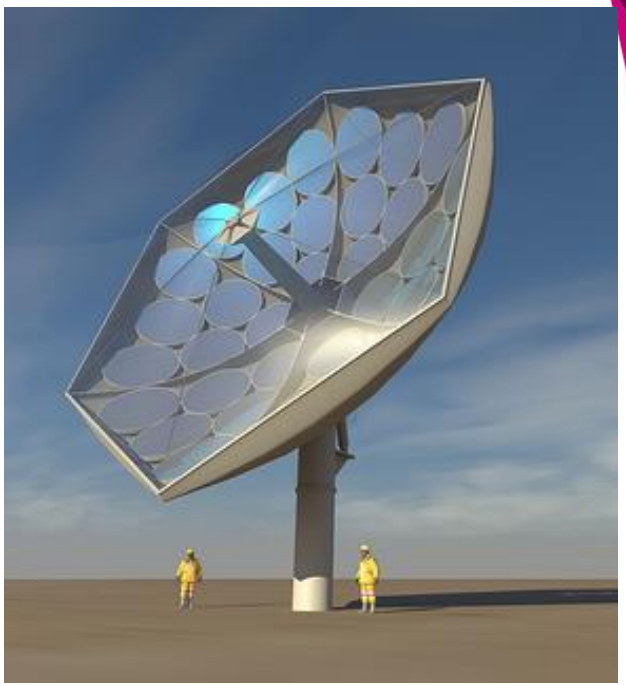


- Heat value vs. temperature and technology
  - *60 ºC datacenter heat: 50% value for space heating*
  - *Desalination and refrigeration: 30 and 5% value*
  - *Technology investments for better reusability (dashed)*
  - *Conversion to electricity below Carnot efficiency (black)*

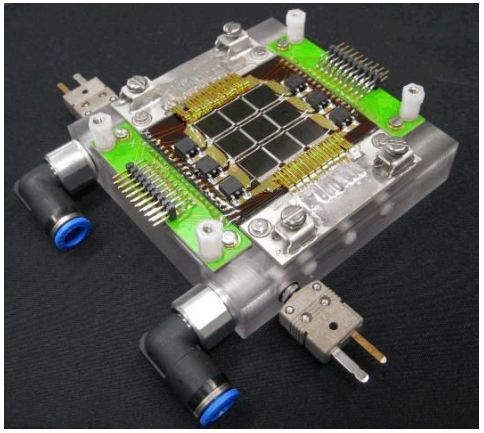


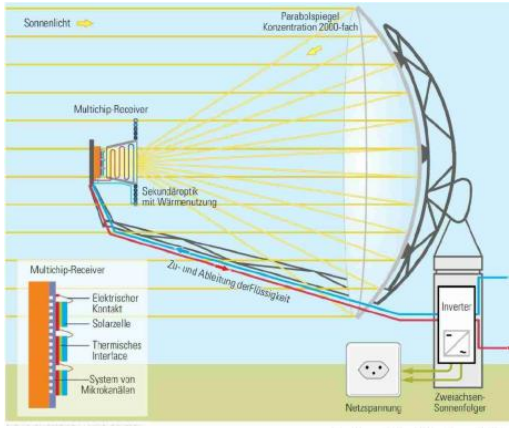
Space heating (blue), Desalination (green), Sorption Cooling (red) ➔

Application of Cooling Technology in Solar Concentrators

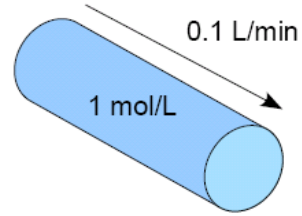


- Swiss KTI funded Project IBM Research, ETH Zurich, NTB Buchs, Airlight Energy Biasca: Low Cost Photovoltaic Thermal Concentrator from innovative materials
- Timeline: 3 Years until commercial prototype
- Size: 25kW electrical and 50 kW thermal @ 90ºC
- Yields: 25% electrical, 50% thermal, and 80% total
- Cost: 250$/$m^2$ aperture (~1$/Wpeak)
- LCOE: <0.1$/KWh for sunny locations
- Microchannel cooled multichip receiver with 10x lower thermal resistance
- Key Aspects:
  Concrete tracking and supporting structure, inflatable mirrors with 10x lower base cost than steel/glass technologies
- Combination with adsorption cooling and membrane distillation desalination (matching interface)
- Extensive economic studies on inclusion of heat-reuse in overall business model
- Free cooling and cooling, and desalination base cases studied
- Sensitivity studies available
- Business Model with assembly of system at deployment site





Uninterruptable Power Supply (UPS) and Chip Power Delivery



Pumped electrolyte solution

- 1 e^- per electroactive species
- full conversion

→ $I \approx$ 160 A

- voltage window 1 V

→ $P \approx$ 160 W

- 0.1 l/min 1 M electrolyte flow carries 160 A (160 W at 1 Volt)
- Full dissipation results in <10ºC heating of 1 Molar solution
- ➔ **Congruent demands for power delivery and cooling!**
- Power delivery now requires many down and up-transformations and DC/DC conversions.
- Overall power loss more than 50%
- Power supply requires a lot of valuable space in vicinity of processors
- Electrochemical power-supply allows one down-conversion to low voltage
- Transport via electrochemical fluid requires less pumping energy than joule heating in copper wires (tubes down to 50 micron)