



IBM Systems Storage

Storage Trends & Futures

Technologieausblick

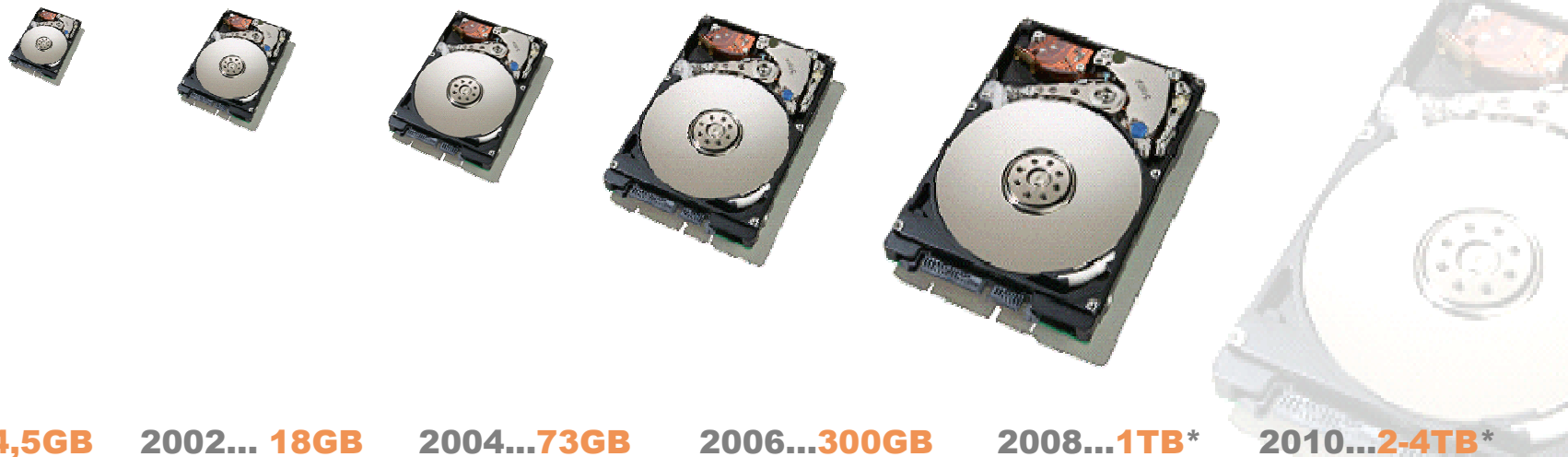
Dr. Axel Koester
Technologist, Storage Consultant
axel.koester@de.ibm.com

Stand der Plattentechnologie

Speicherklasse 'Solid State Memory'

Die Zukunft von Speicher & SAN

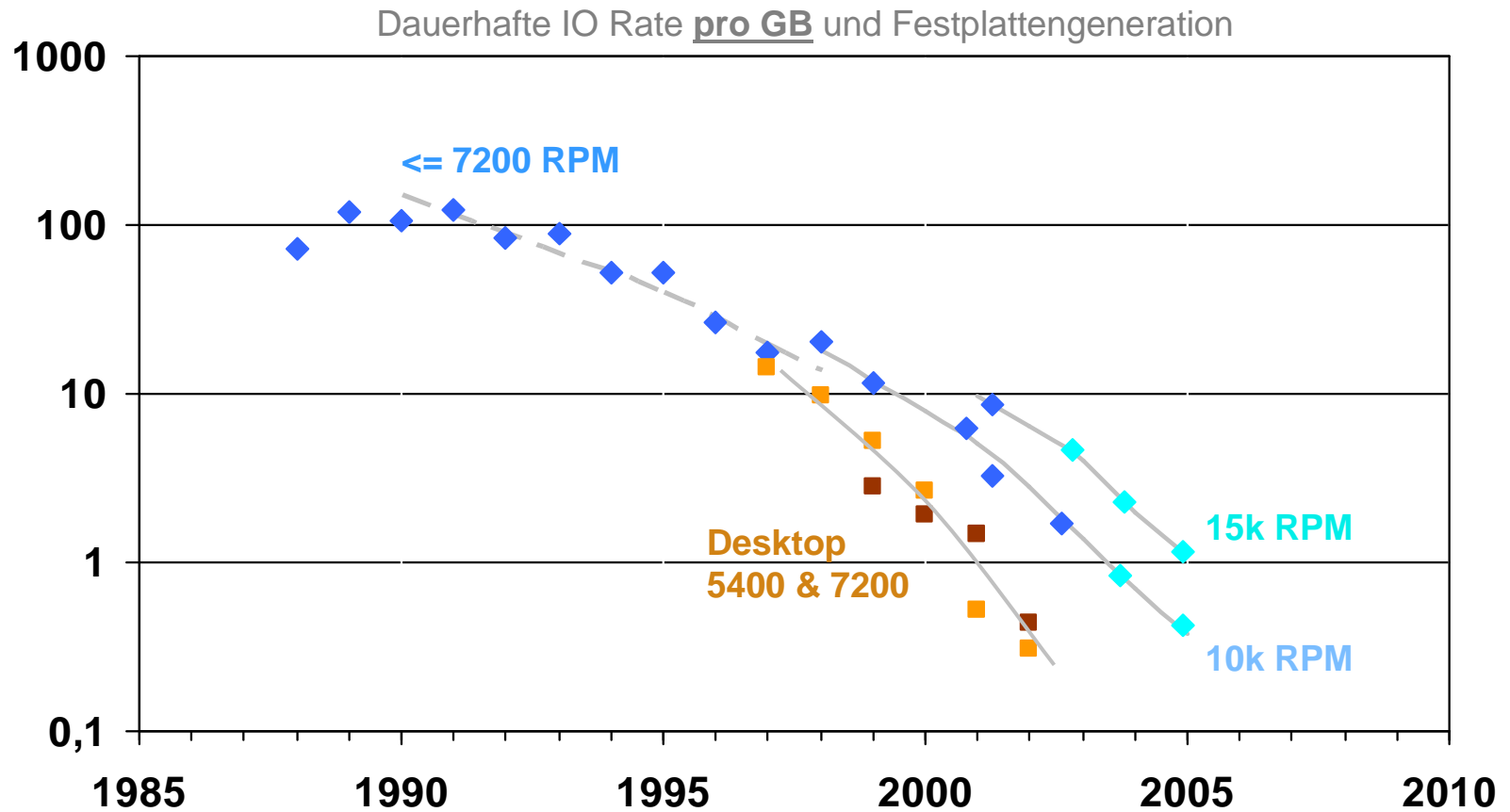
Wachstum der Festplatten



Festplatten wurden im Jahrzehnt 1000 x größer, aber nicht gleichermaßen schneller!

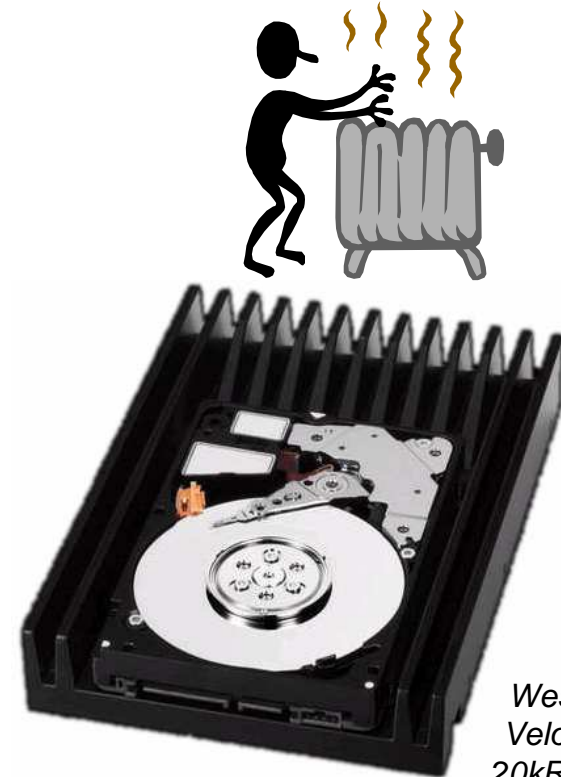
*SATA

Festplatten-Datenrate *pro Gigabyte* fällt dramatisch



Schnellere Festplatten? 20k RPM?

- 20.000 RPM Platten laufen sehr **warm**
- RPM $\times 2$ = **Leistungsaufnahme $\times 8$**



Western Digital®
VelociRaptor 2.5"
20kRPM Prototype

- → kleinerer Scheiben-Ø
- → größerer Motor-Ø
- → höherer Preis/GB = Widerspruch

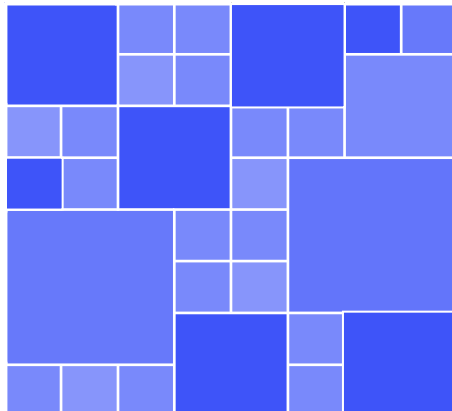
(*) Luftreibungsverluste $\sim \{\text{RPM}\}^3$

Schnellere Antwortzeiten durch Cache Innovation

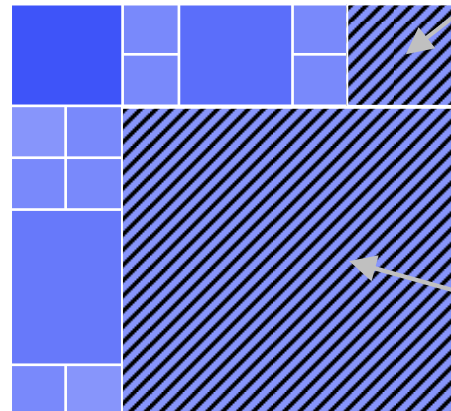


Datenbanken und Transaktionssysteme

Cache Innovation SARC (ursprüngliche Publikation 2003)



DS8000 Lesecache
im 100% full Status



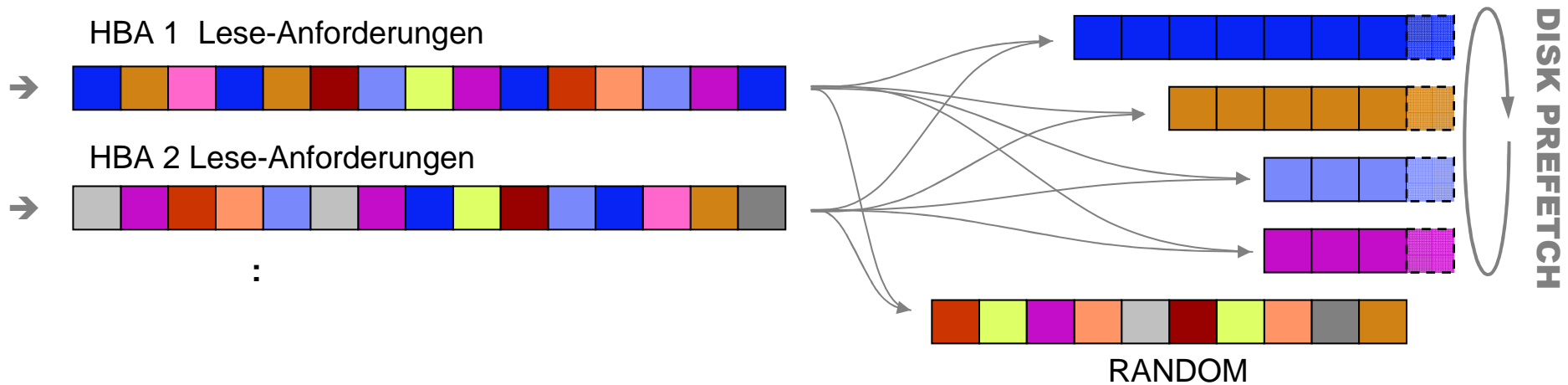
DS8000 Cache nach
dem Lesen eines
großen Backups

- Standard Algorithmus **LRU** (*least recently used*) entfernt den ältesten Eintrag
- IBM Algorithmus **SARC*** entfernt den voraussichtlich nicht mehr benötigten Eintrag
- Balance zwischen LRU und LFU (**wie alt** vs. **wie beliebt**)

* SARC: Simplified Adaptive Replacement Cache, LFU: least frequently used

Effektive Cachegröße: +33%

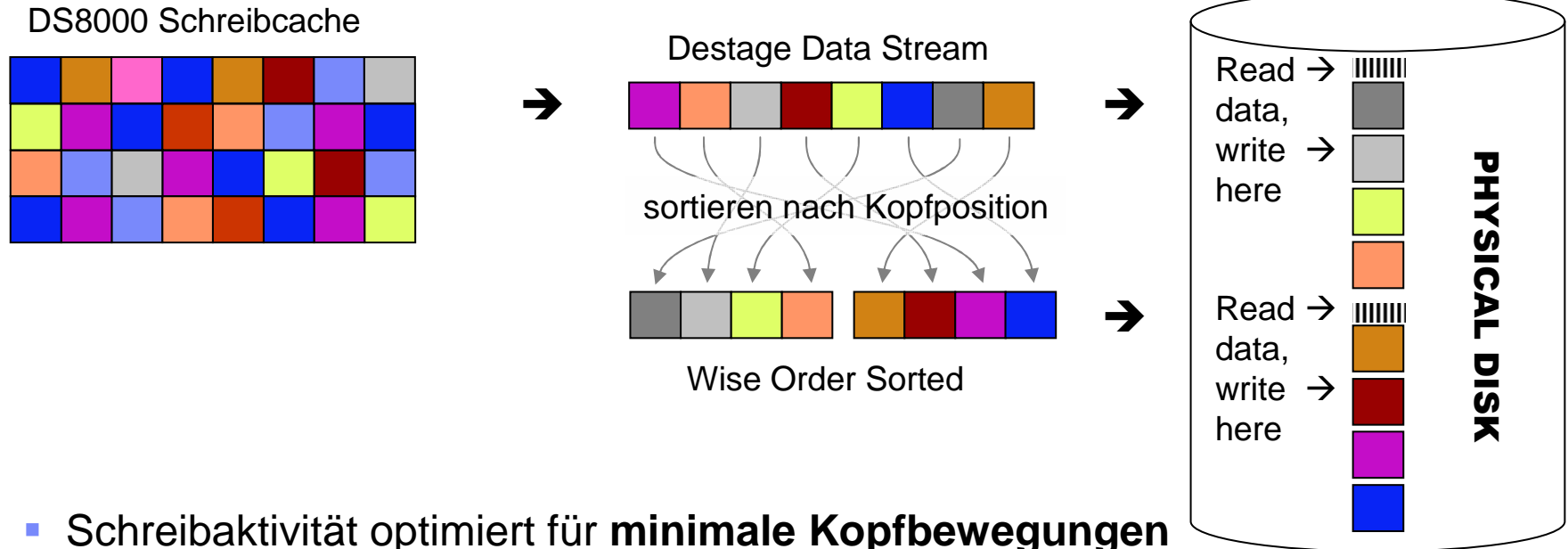
DS8000 Cache: Adaptive Multistream Prefetch



- Adaptive Multistream Prefetch AMP findet in chaotischen Datenmustern zusammengehörige sequentielle Zugriffe
- über alle Adapter und IO clients hinweg
- in Echtzeit bei > 120.000 IO/s

DS8000 Intelligent Write Caching (*brandneu*)

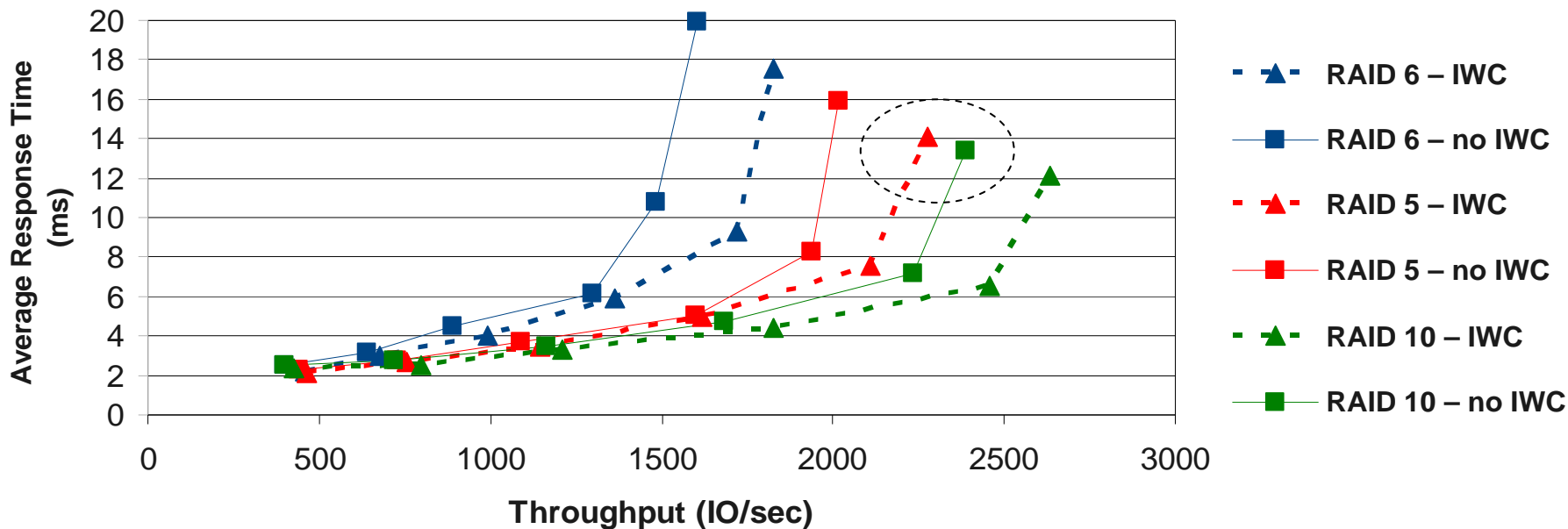
"Wise Order Writes"



- Schreibaktivität optimiert für **minimale Kopfbewegungen**
- Ideal bei hohem Verteilungsgrad (große Pools, Striping)
- Verringerter Schreib-Overhead durch Ausnutzung der aktuellen Kopfposition

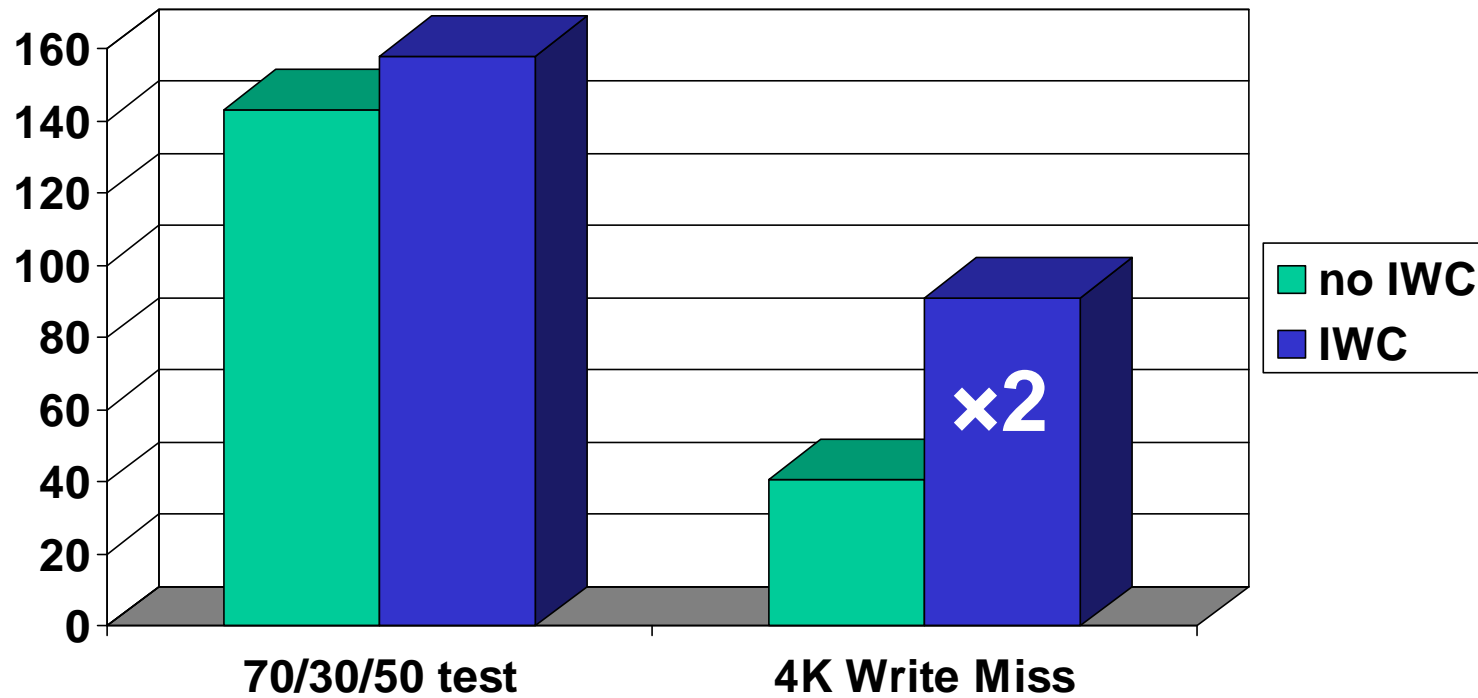
Intelligent Write Caching – Die Praxis

**DB Open Performance
70/30/50 Load
(single 8 disk arrays)
15 K RPM Disks**



RAID 5 bei 70/30/50 Workload nun fast so schnell wie bisheriges RAID 10.
RAID 10 wird abermals um 10% schneller, RAID 6 deutlich effizienter.

Intelligent Write Caching – für Datenbanken

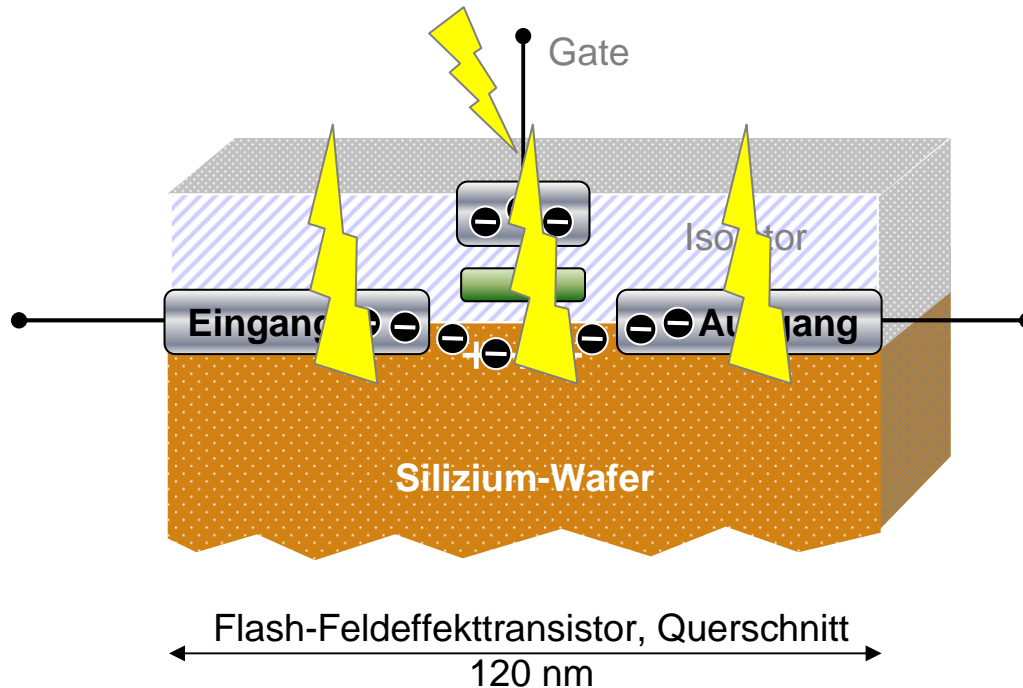


RAID 10, 15K RPM, mix of 146, 300 and 450 GB
 (64) (16) (16)

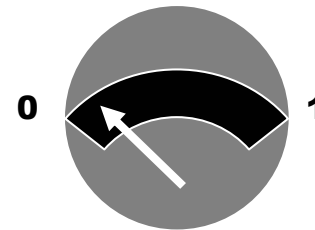
Disk versus(?) Flash



Funktionsweise von Flash-Speicher



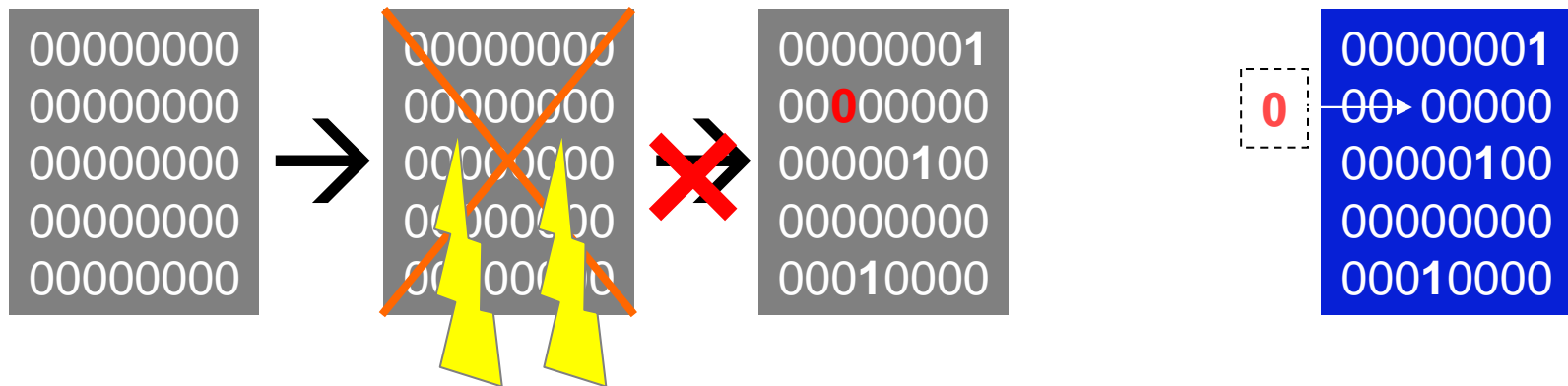
- **Floating Gate** (isoliert)



- Es können nur **EINSEN*** geschrieben werden
- Individuelles Löschen nicht möglich, nur Blocklöschung
- Löschen = Alterung

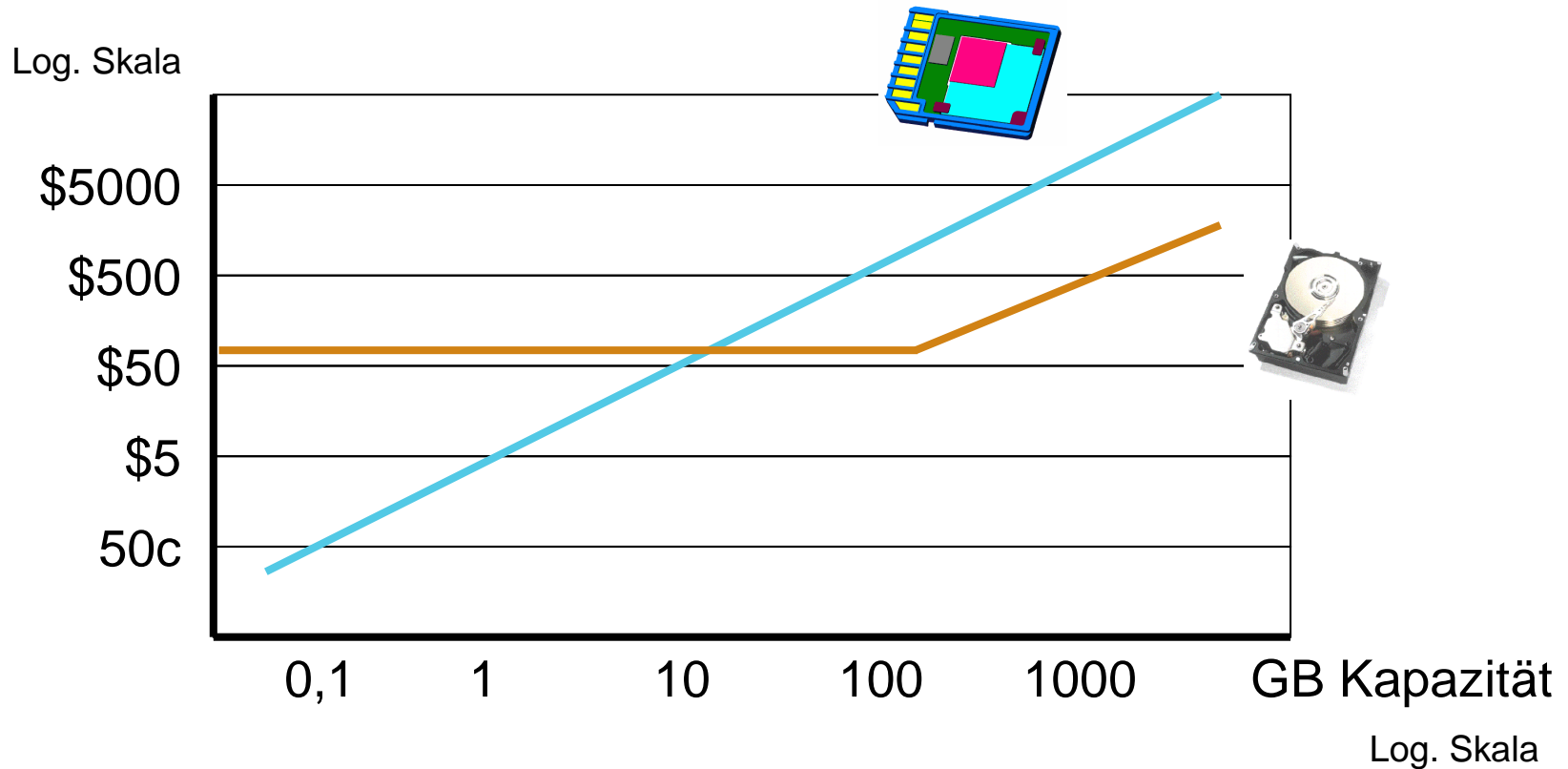
* Bei NAND Flash ist der aktive Zustand *NULL*, in diesem Fall sinngemäß "EINSEN" durch "NULLEN" ersetzen

Funktionsprinzip NAND-Flash : wie EEPROM



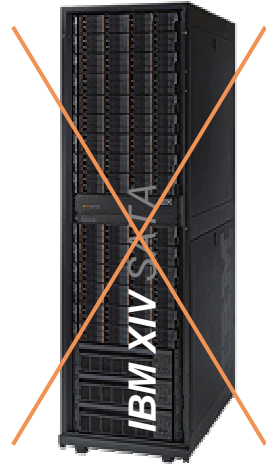
- Für **jede Schreiboperation** ist eine Blockkopie notwendig
- **Random Write** ist nicht der optimale Workload für NAND Flash
- Zuvor genutzter Block bleibt bis zum Löschmodus **gesperrt**
- 10^5 Löschvorgänge je Block möglich, optimal streuen!

Preisentwicklung Flash versus Disk nach Kapazität



Nicht maßstabsgetreu; 2008 Schätzwerte

Solid State Disks SSD im IBM Portfolio



DS8000

**73/146 GB SSD
3,5" Carrier**



Bladecenter

**64 GB SSD
internal**



DS5000

**73/146 GB SSD
3,5" Carrier**



**SAN Volume
Controller**

**Diverse Optionen
(Ankündigung demnächst)**

IBM SVC QUICKSILVER : 1 Mio echte IOPS

- Technologie-Demonstrator 2008:
IBM SAN Volume Controller + integrierter Flashspeicher
- Datenbank 70/30 Workload, 0% Cache Hit, Laufzeit 2 Stunden,
1 Mio IOPS gemittelt bei **700µs Antwortzeit**

SVC Controller →
(SVC Code 4.3)



Flash SVC Controller →
(modif. SVC Code 4.3)

Nicht maßstabstreue
Prinzipdarstellung !

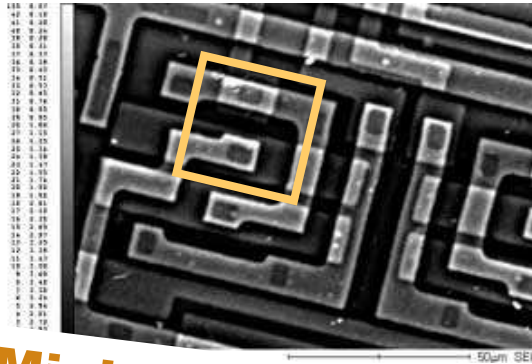
Alle Komponenten
aus dem IBM
Standardportfolio.

Flash von FusionIO®

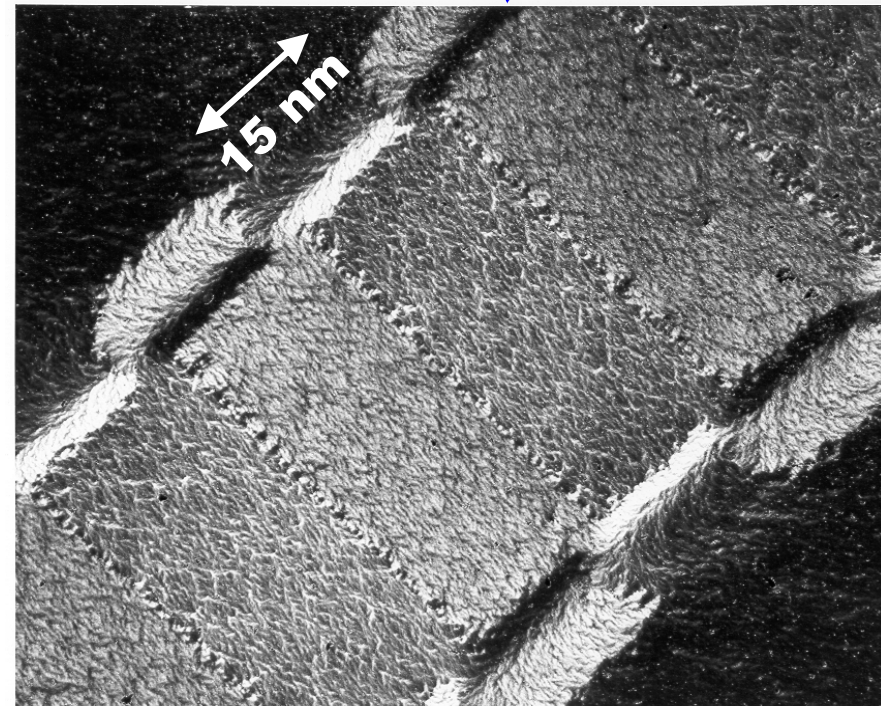
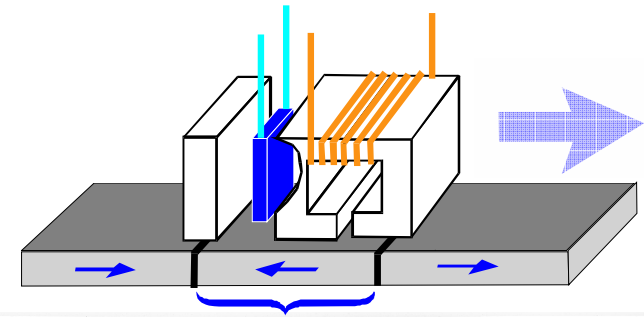
Chip- oder Analoogspeicher?



Chip-Speicher oder magnetischer Speicher ?



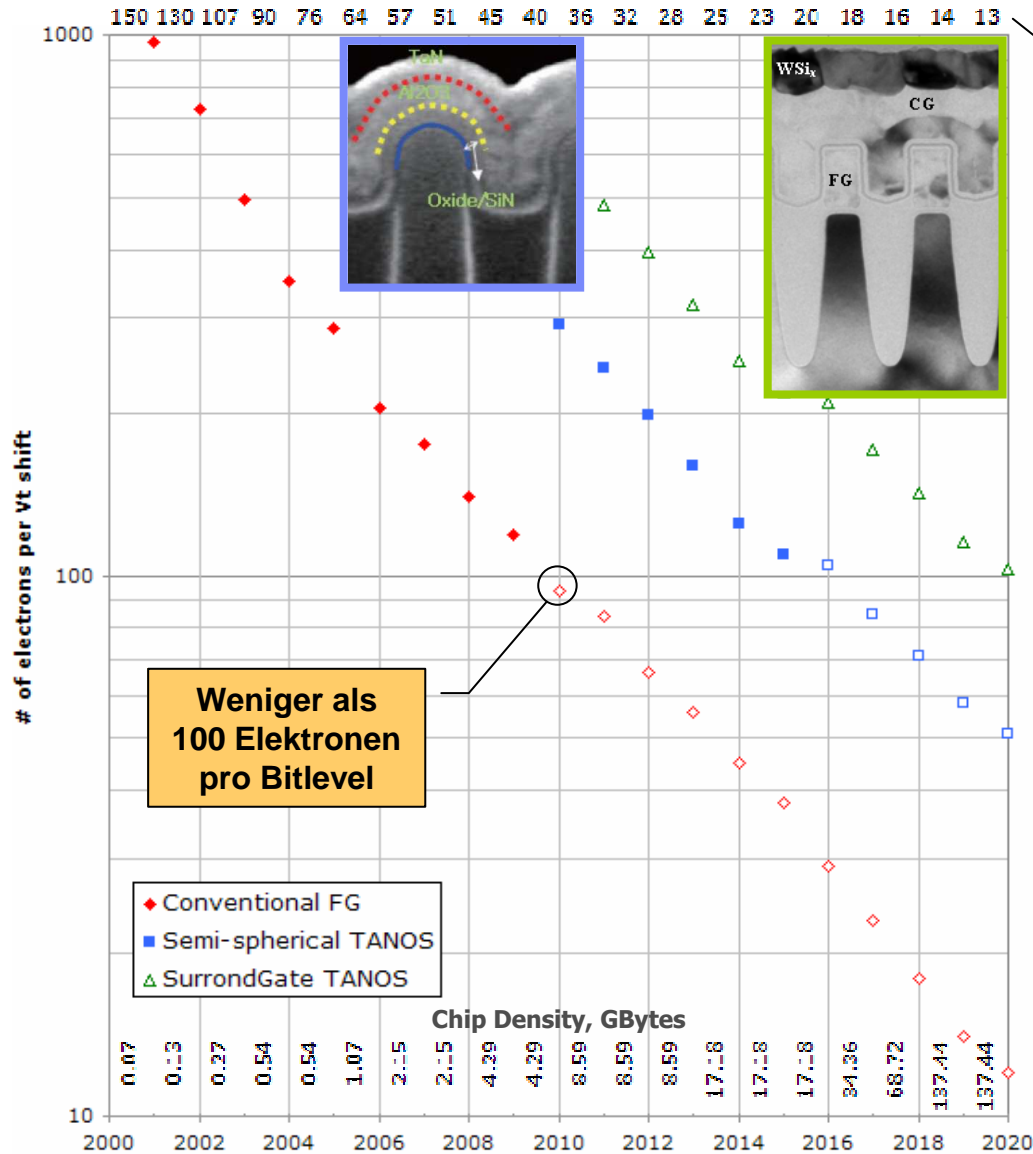
**Minimum je Bit
ca. 40x40 nm**



Disk :

Chip-Speicher werden voraussichtlich noch geraume Zeit der Speicherdichte (und dem Preis) großer Festplatten nacheilen.

Chiptechnologie bei zunehmend feiner Lithographie



Lithografie-Auflösung (Prognose)

Kaum mehr als 100 Elektronen bestimmen heute ein Bit (MLC).

Pro Monat darf nicht einmal **ein Elektron** verloren gehen.

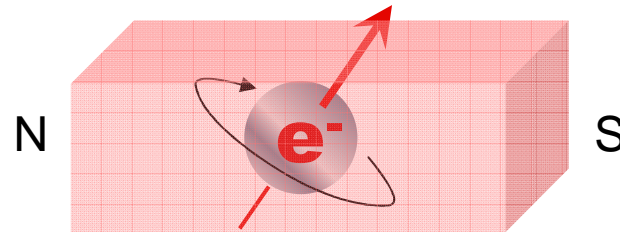
→ **Verschleiß-Anfälligkeit** steigt

Ausweg: Mehr Elektronen trotz noch kleinerer Strukturen durch **sphärische** oder **zirkuläre** Gates (TaNOS = Tantal-Nitrid-Oxid-)

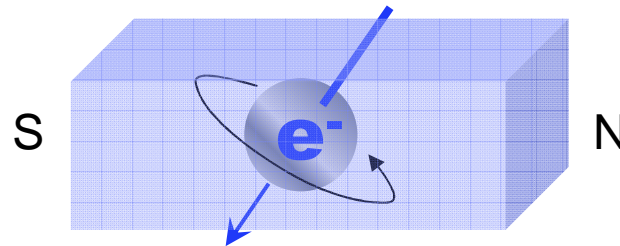
Weniger als 100 Elektronen pro Bitlevel

Quelle: IBM / Samsung

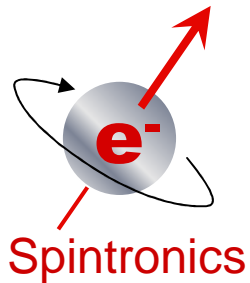
Spintronics: Neue Klasse von Festkörperspeichern



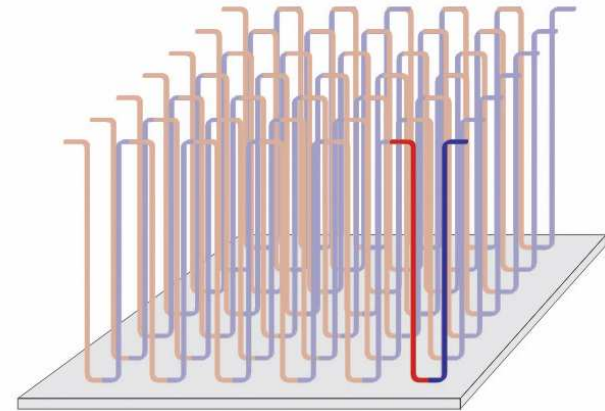
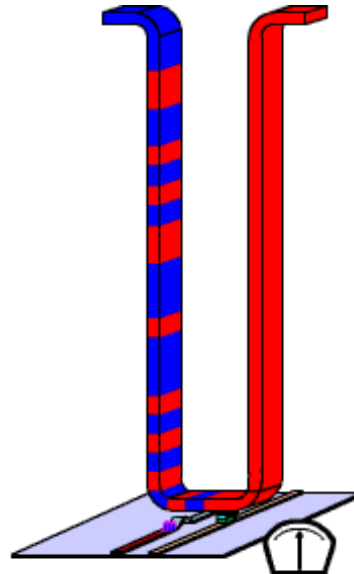
Spintronics



Speichern in 3D : "RaceTrack" für Spin-Elektronen

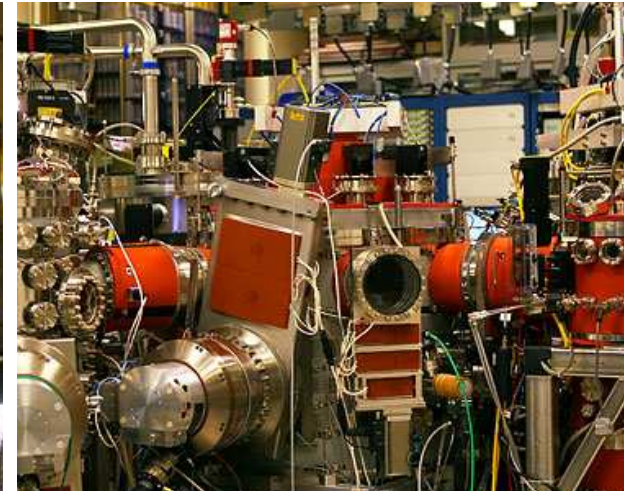


"Großer" Schreib/Lese-Kopf,
"kleine" Spin-Datenfelder auf
ferroelektrischem Nanodraht

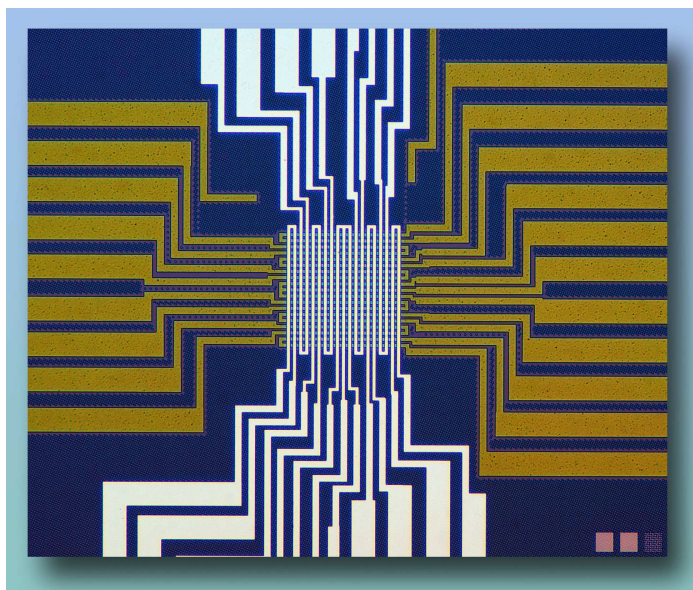


RaceTrack Storage Array:
Hohe Datendichte in 3D

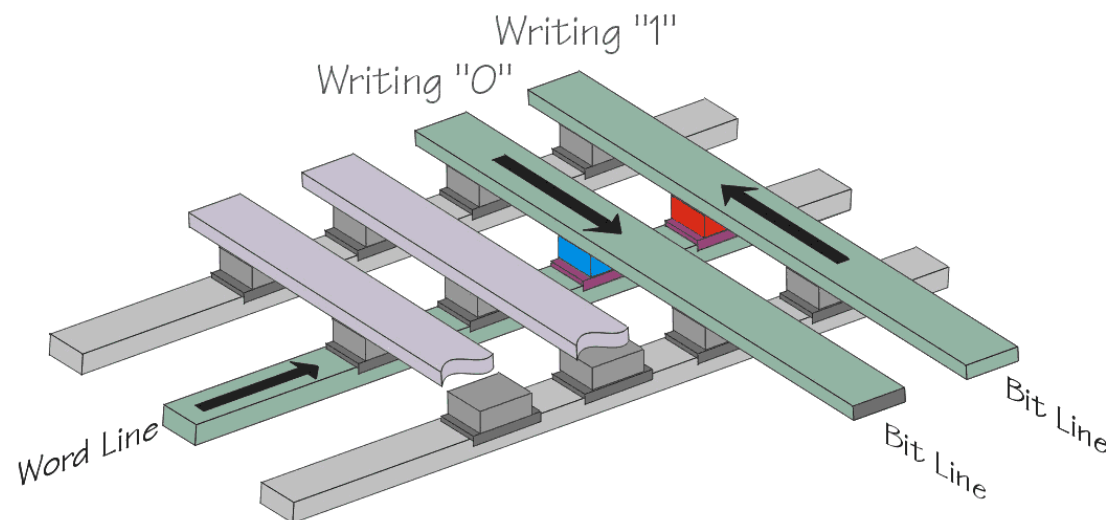
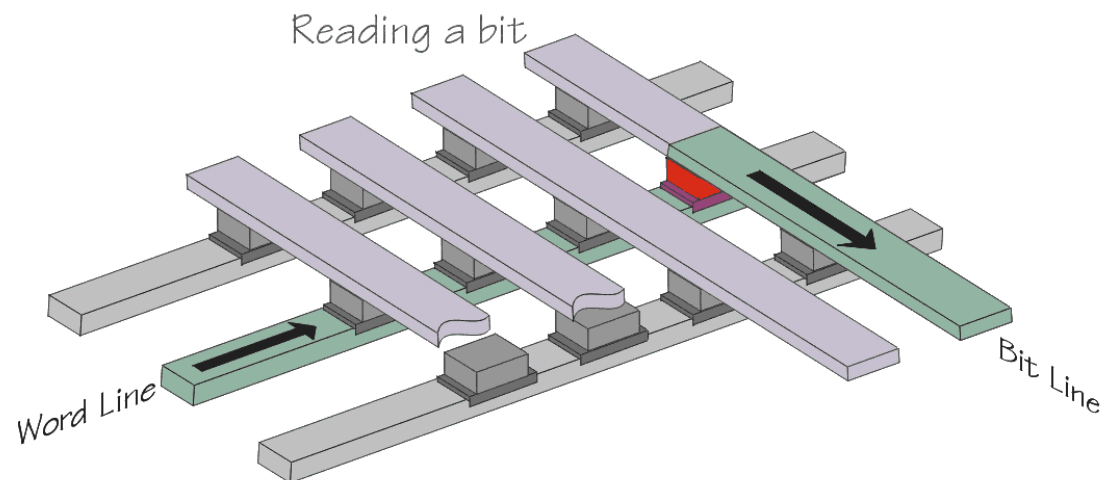
IBM Fellow Stuart Parkin,
Erfinder der GMR Leseköpfe,
erforscht "Racetrack Memory"



Magnetisches RAM

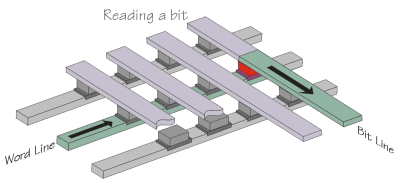
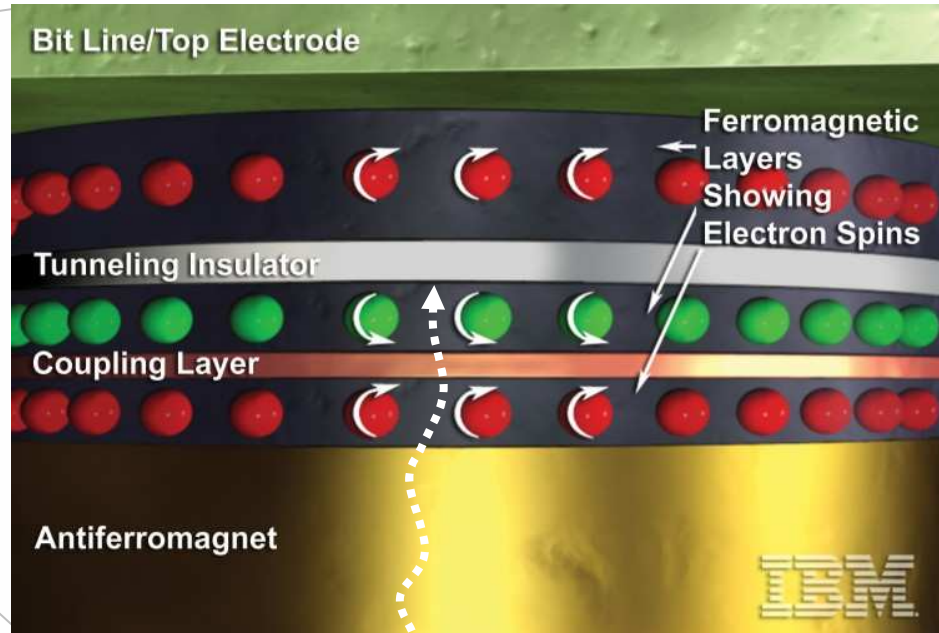
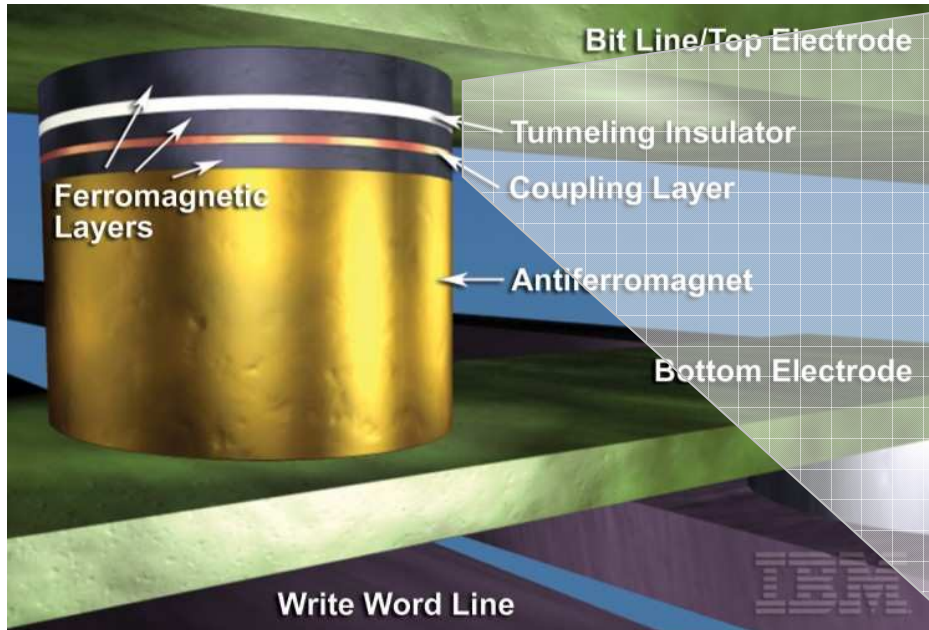


IBM Prototyp 199..



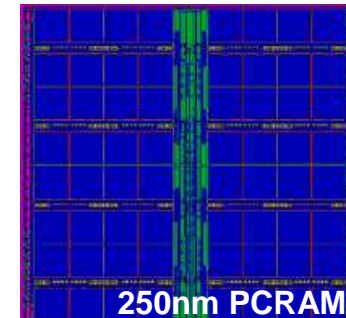
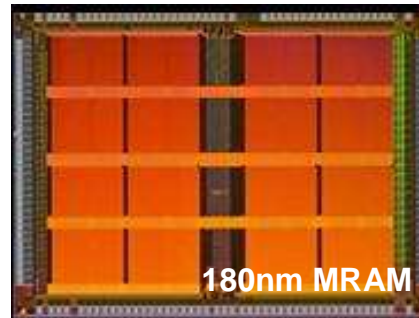
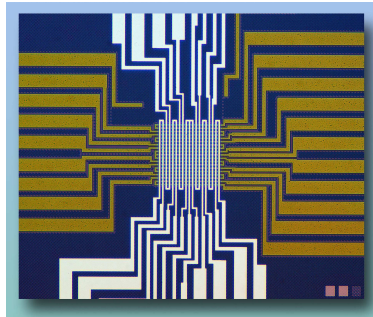
Bsp. Hersteller: Freescale Inc.
"MR2A16A" 4Mb nichtflüchtig
@ 35 nsec Zugriffszeit.

Magnetisches RAM



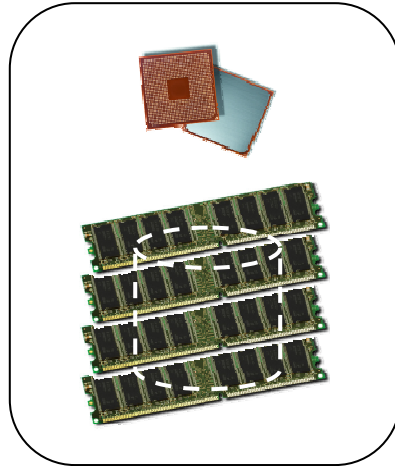
**Neuer IBM Demonstrator:
2 nsec Zugriffszeit**
(10 mal schneller als heutiges DRAM)

Nicht-flüchtiges RAM = IT Revolution !

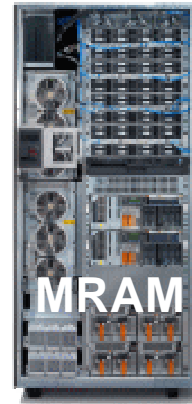


- Magnetic RAM / Ferroelectric RAM / Phase-Change RAM
- Flash Memory = Übergangstechnologie ("marktbereitend")

Wie verbaut man *schnellen* Solid-State Speicher?



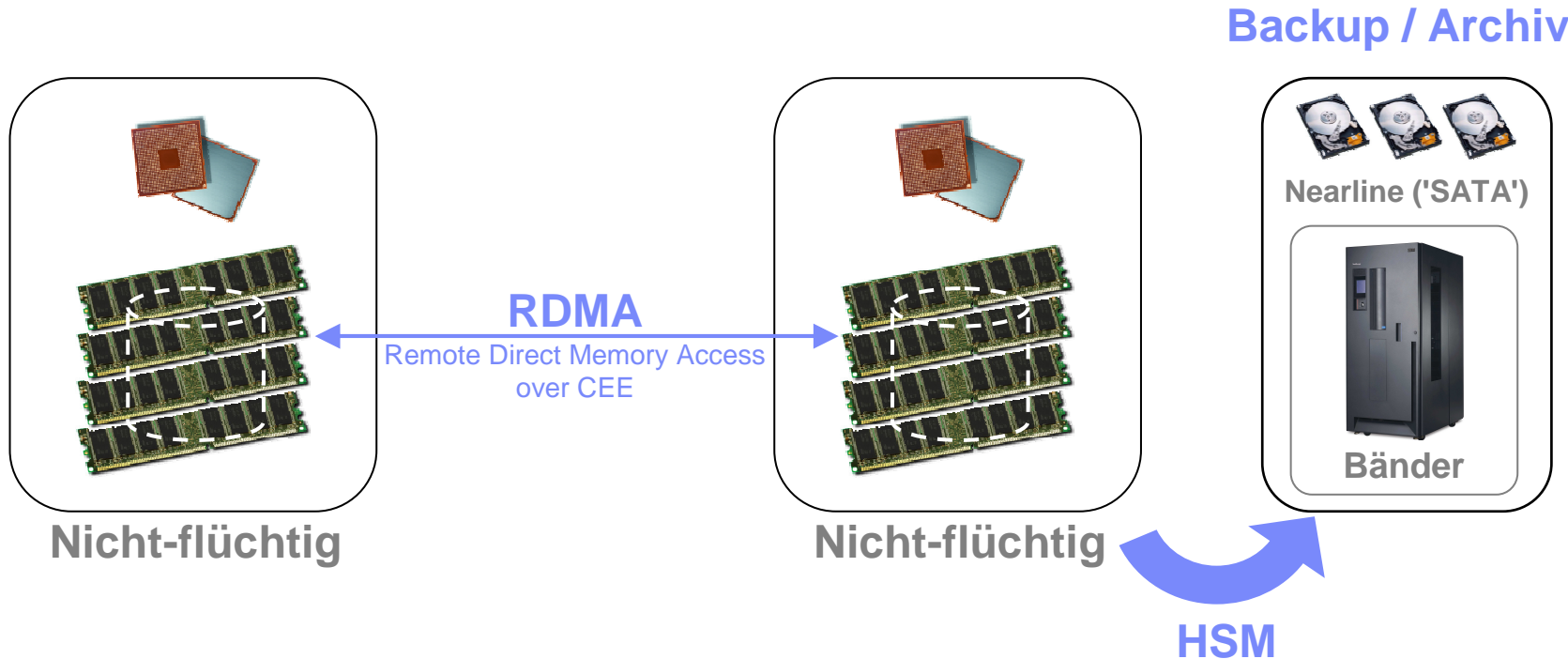
20 nsec



2 nsec

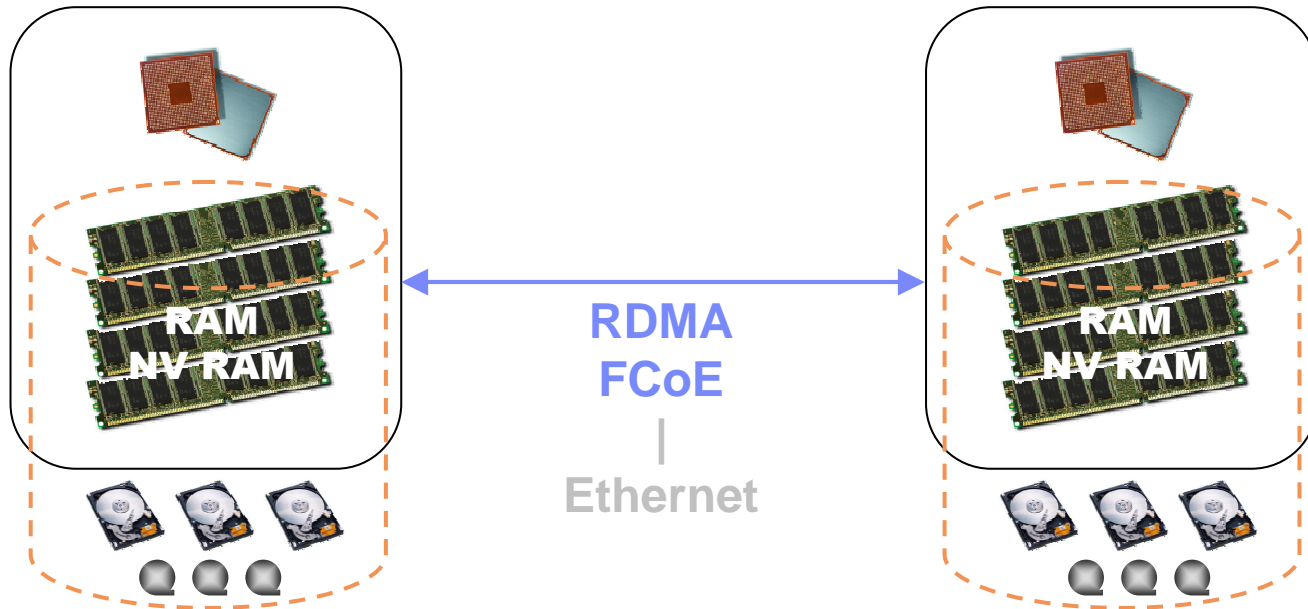
- Online Speicher zieht zum Prozessor
- Schnelle (Fibrechannel-) Platten verschwinden
- Monolithisches Design = hohe Verfügbarkeit

Die SANs der Zukunft



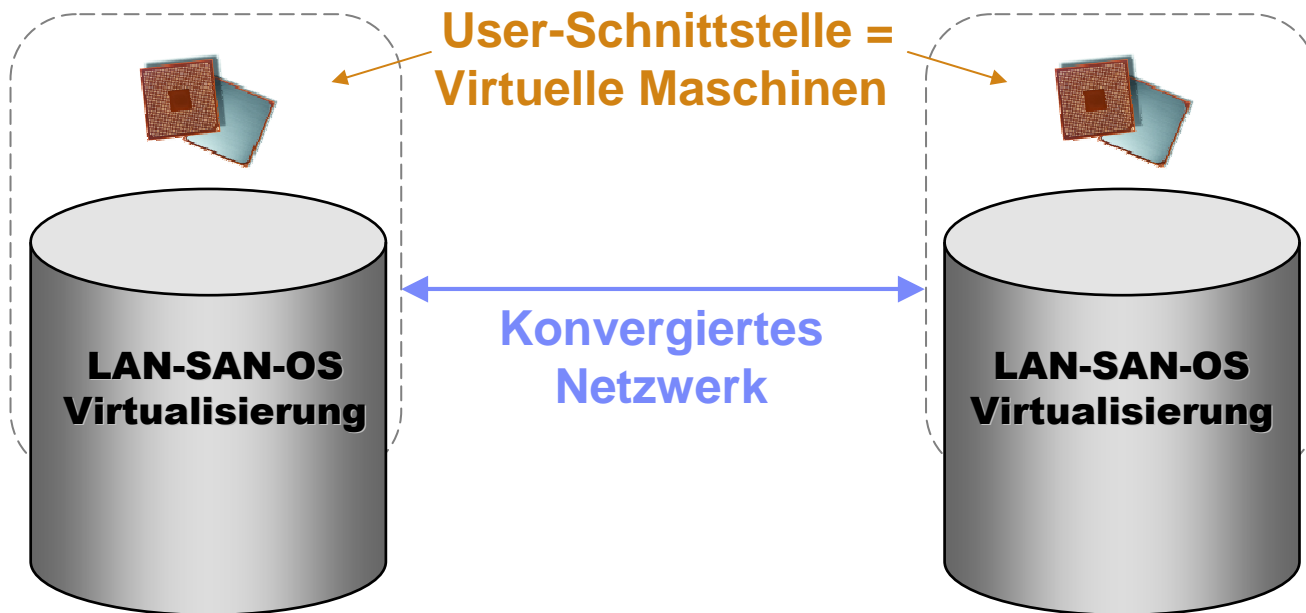
- **Memory-to-memory SAN** (*RDMA über Converged Enhanced Ethernet*)
- Platten → RAM
- Bänder → Platten
- Paging → HSM

Storage in der Zukunft : *Multi-Tier* Speicher



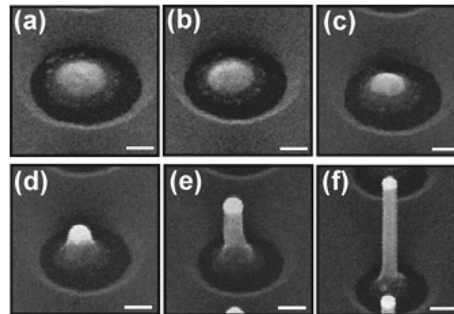
- Das konvergierte Netzwerk verknüpft *alle* Speicherklassen von RAM bis Tape, nebst Nachrichten- und Dateiübertragung.
- Applikationen sehen nur noch RAM und RAM LUNs (zwecks Kompatibilität). Hierarchieverwaltung verläuft davon getrennt.

VM Speicher/Server

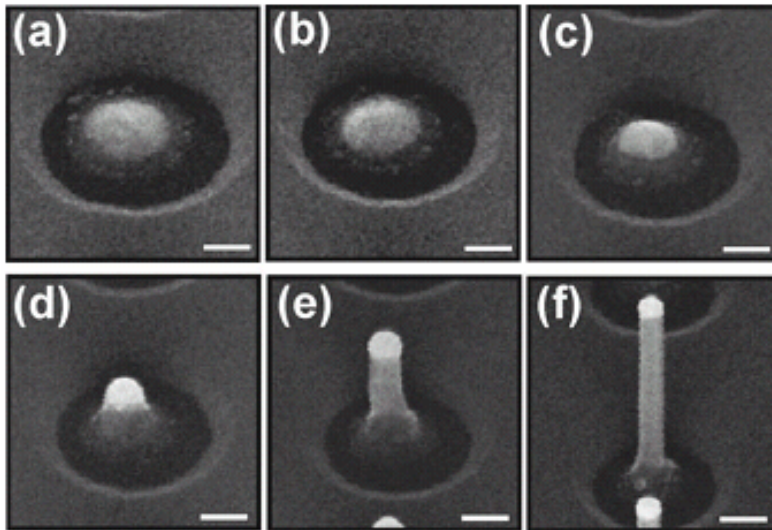


- Regelbasiert
- Komplexität wird vom Hersteller gehandhabt
- Software hat keine Systemsicht mehr

Die Nano-Zukunft

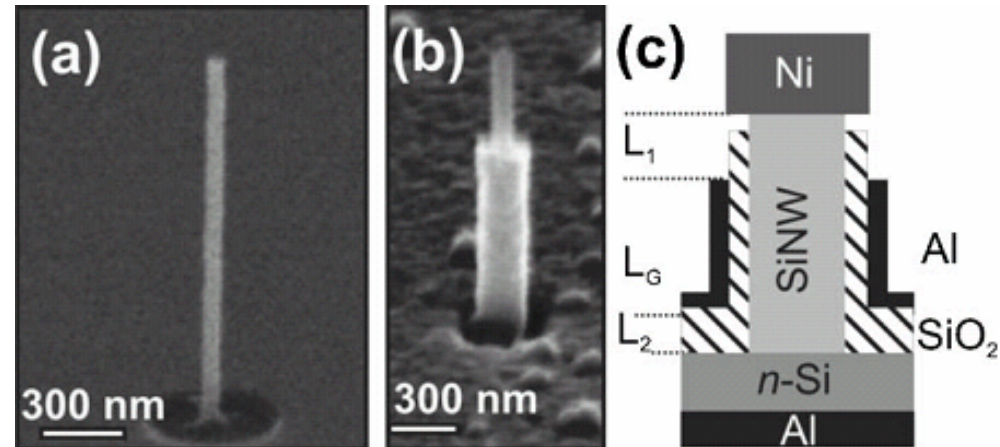


Die schnellsten & sparsamsten Transistoren



← Wachstum von Silizium-Nanosäulen

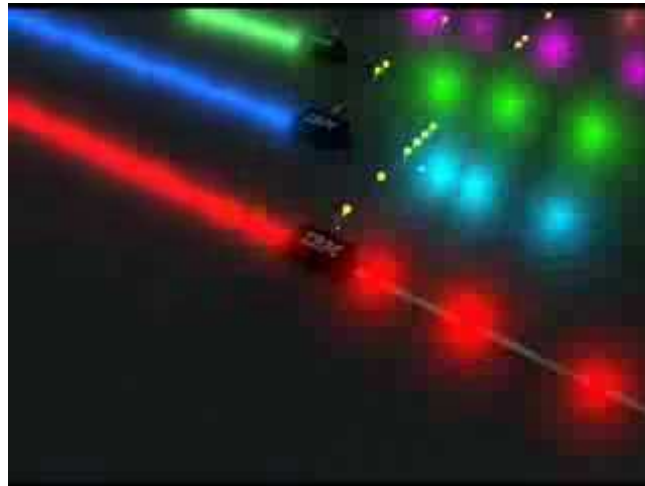
Anwendung: Schnell sperrender Transistor



Mit Mantel-Gate versehene Nanosäule

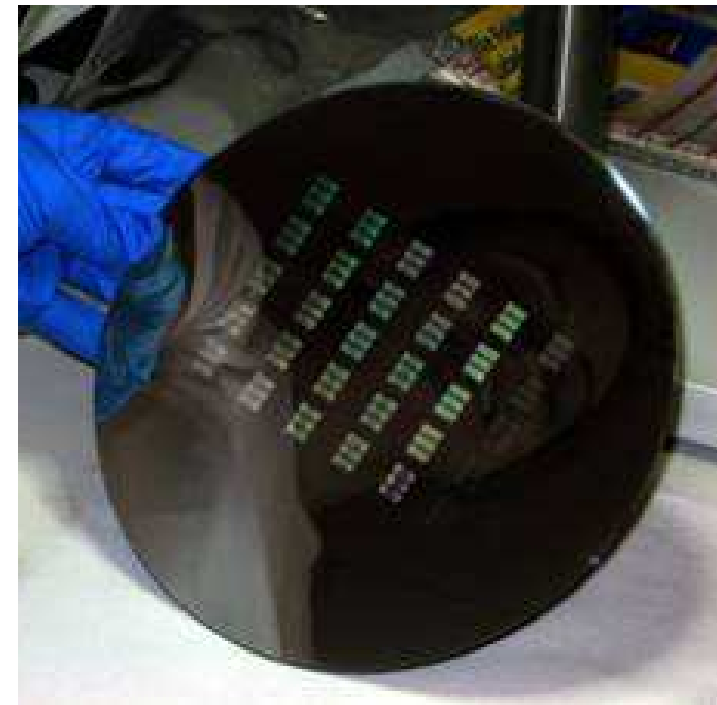
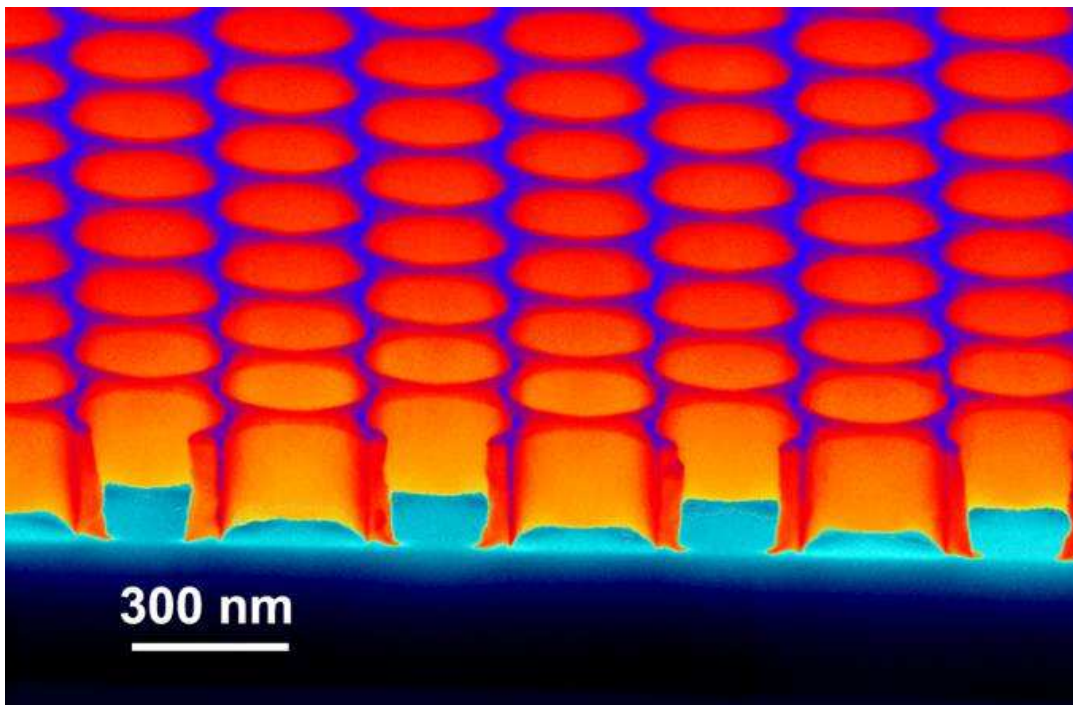
- Minimaler Verbrauch
- Höchste Schaltgeschwindigkeit
- TeraHertz nicht ausgeschlossen

Die optische Zukunft



Langsames Licht

- Nano-Silizium mit **300-fachem** Brechungsindex
- Licht kann **variabel** gebremst werden





axel.koester@de.ibm.com

Disclaimer

No part of this document may be reproduced or transmitted in any form without written permission from IBM Corporation.

Product data has been reviewed for accuracy as of the date of initial publication. Product data is subject to change without notice. This information could include technical inaccuracies or typographical errors. IBM may make improvements and/or changes in the product(s) and/or program(s) at any time without notice. Any statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

The performance data contained herein was obtained in a controlled, isolated environment. Actual results that may be obtained in other operating environments may vary significantly. While IBM has reviewed each item for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customer experiences described herein are based upon information and opinions provided by the customer. The same results may not be obtained by every user.

Reference in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business. Any reference to an IBM Program Product in this document is not intended to state or imply that only that program product may be used. Any functionally equivalent program, that does not infringe IBM's intellectual property rights, may be used instead. It is the user's responsibility to evaluate and verify the operation on any non-IBM product, program or service.

THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER EXPRESS OR IMPLIED. IBM EXPRESSLY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR INFRINGEMENT. IBM shall have no responsibility to update this information. IBM products are warranted according to the terms and conditions of the agreements (e.g. IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided. IBM is not responsible for the performance or interoperability of any non-IBM products discussed herein.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products in connection with this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.