IBM Systems Storage

Storage Trends & Futures
Technologieausblick

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

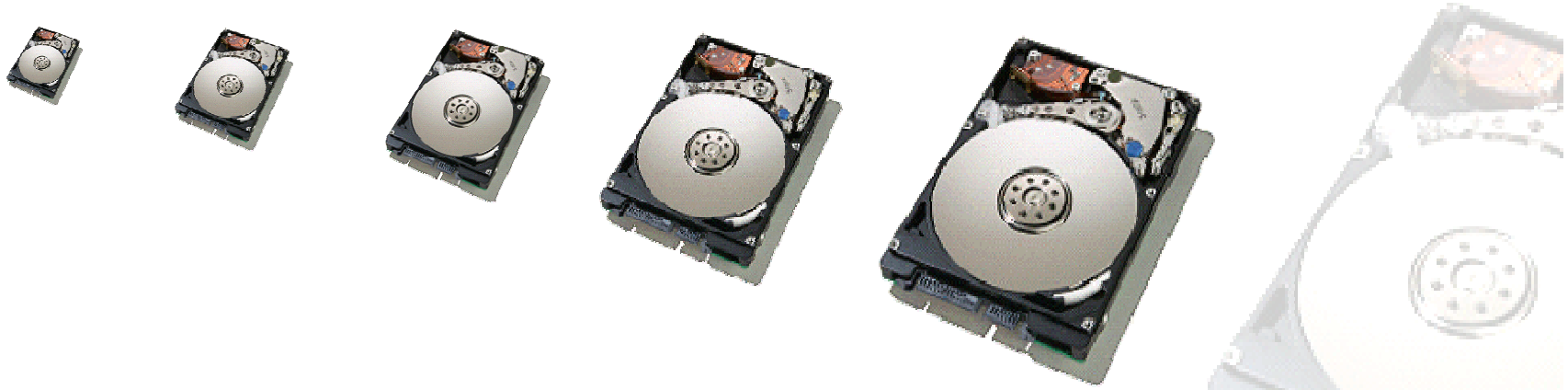
04/2009

Stand der Plattentechnologie

Speicherklasse 'Solid State Memory'

Die Zukunft von Speicher & SAN

Wachstum der Festplatten

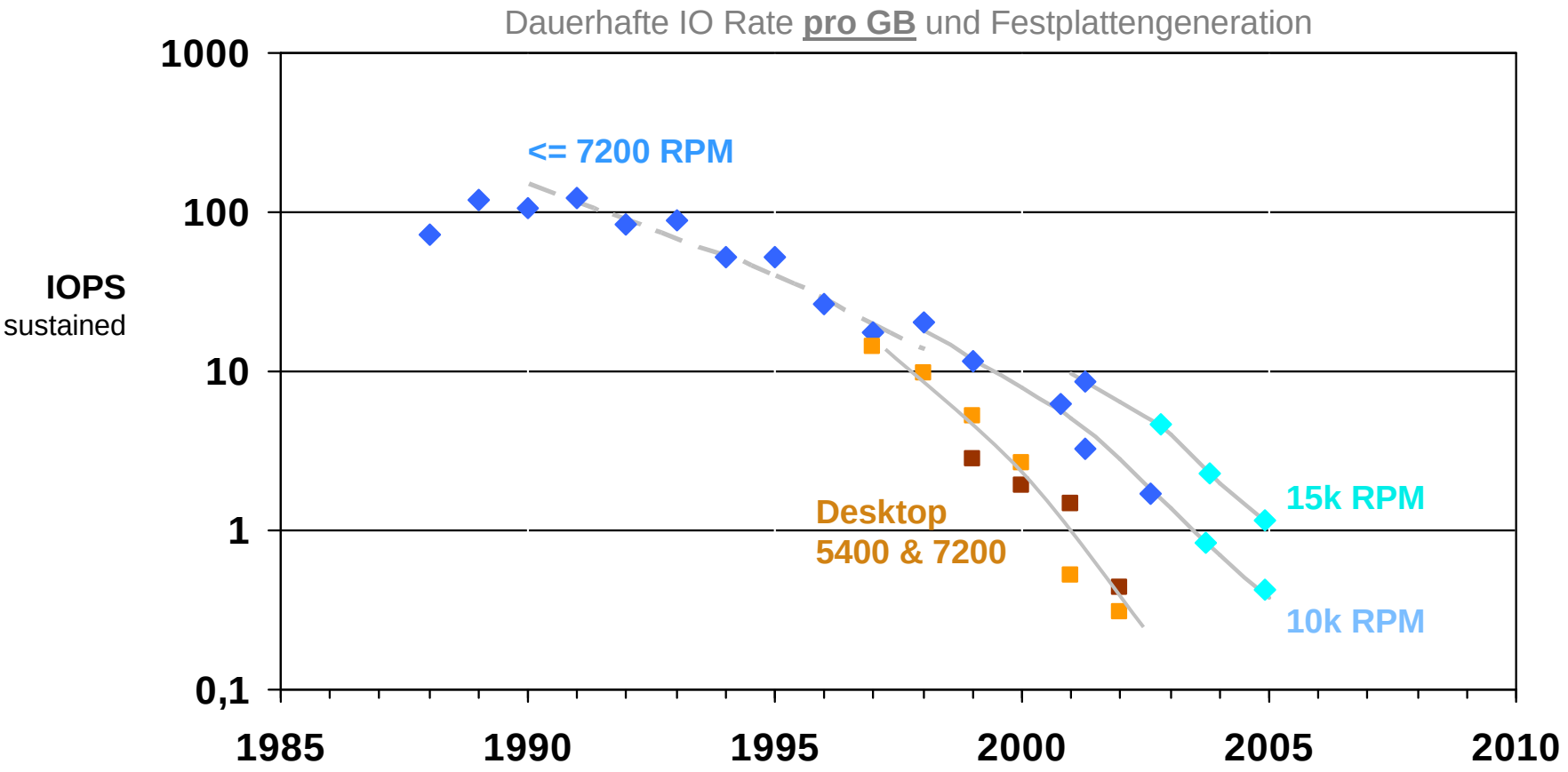


2000... **4,5GB** 2002... **18GB** 2004...**73GB** 2006...**300GB** 2008...**1TB*** 2010...**2-4TB***

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

*SATA

Festplatten-Datenrate *pro Gigabyte* fällt dramatisch



Schnellere Festplatten? 20k RPM?

- 20.000 RPM Platten laufen sehr **warm**
- **RPM ×2 = Leistungsaufnahme ×8**

Western Digital ®
VelociRaptor 2.5"
20kRPM Prototype

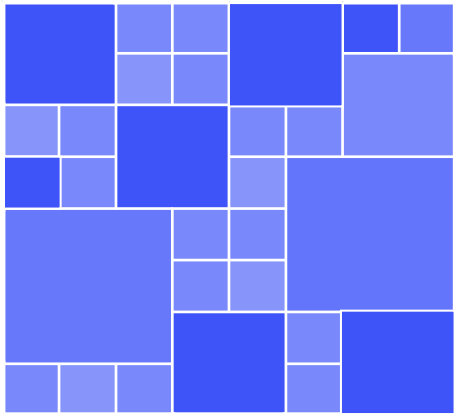
- → kleinerer Scheiben-Ø
- → größerer Motor-Ø
- → höherer Preis/GB = Widerspruch

(*) Luftreibungsverluste ~ {RPM} 3

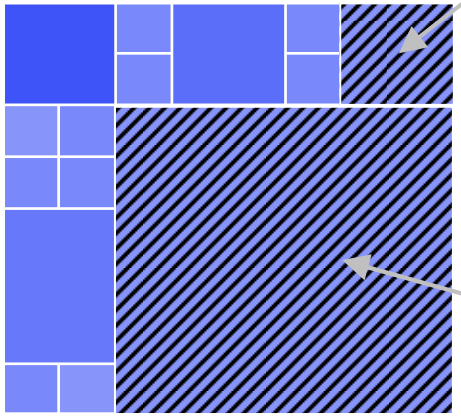
Schnellere Antwortzeiten durch Cache Innovation

Datenbanken und Transaktionssysteme

Cache Innovation SARC (ursprüngliche Publikation 2003)



DS8000 Lesecache im 100% full Status



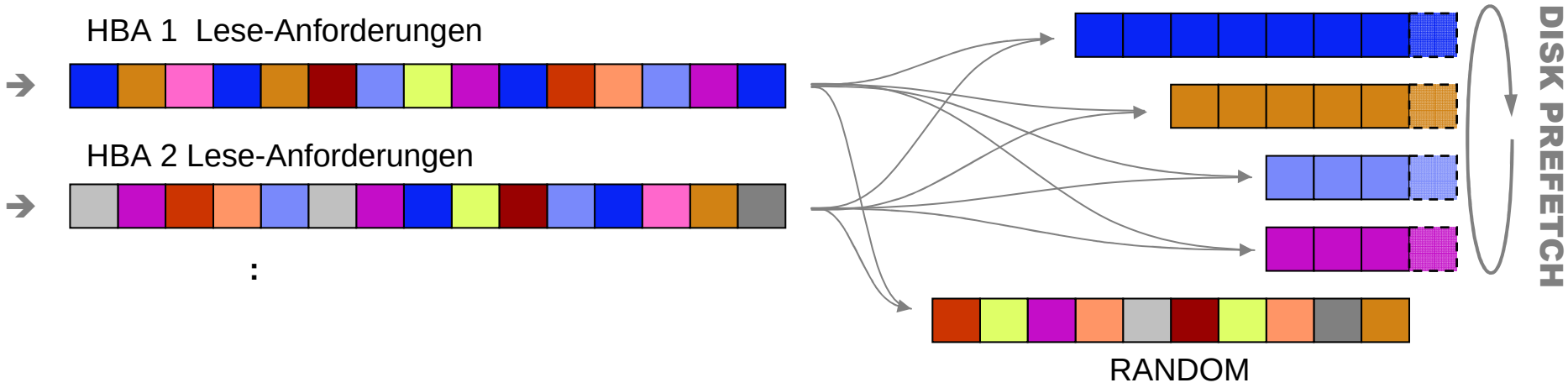
DS8000 Cache nach dem Lesen eines großen Backups

- 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%

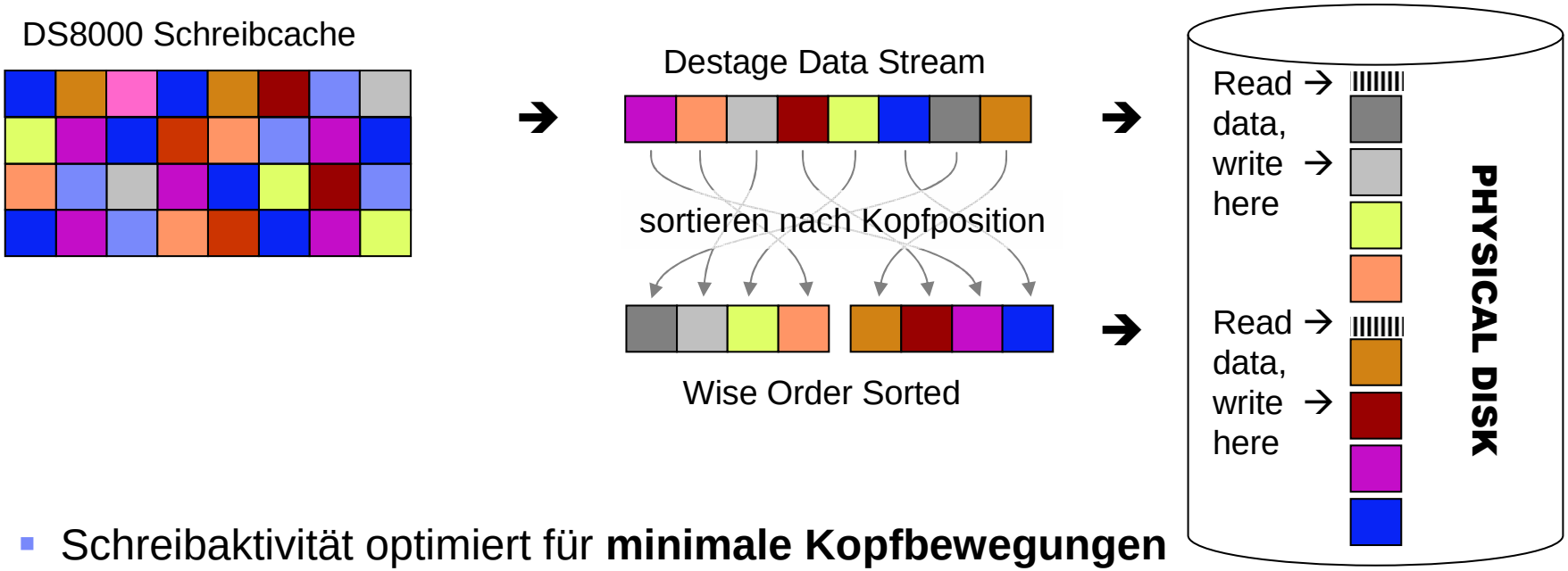
DS8000 Cache: Adaptive Multistream Prefetch



- 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

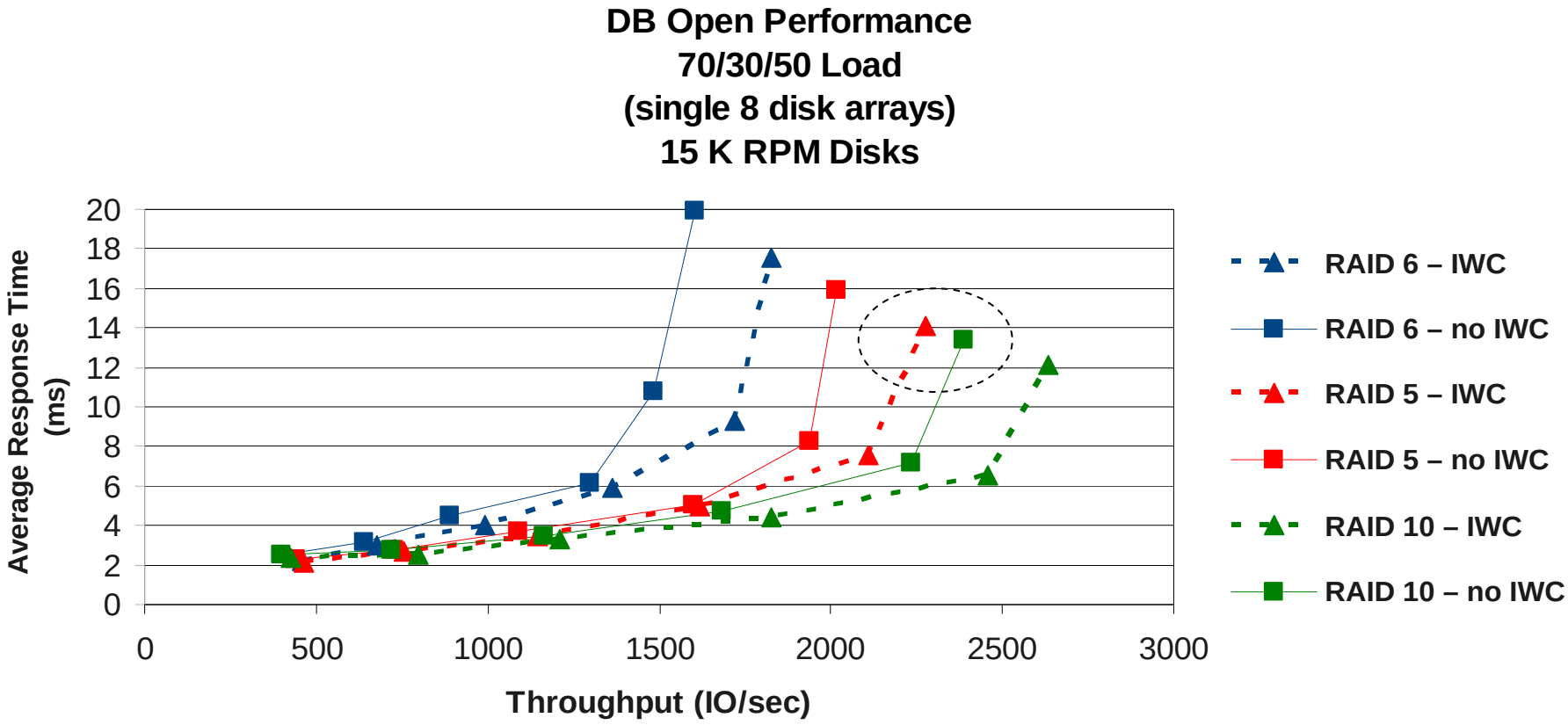
DS8000 Intelligent Write Caching *(brandneu)*

"Wise Order Writes"



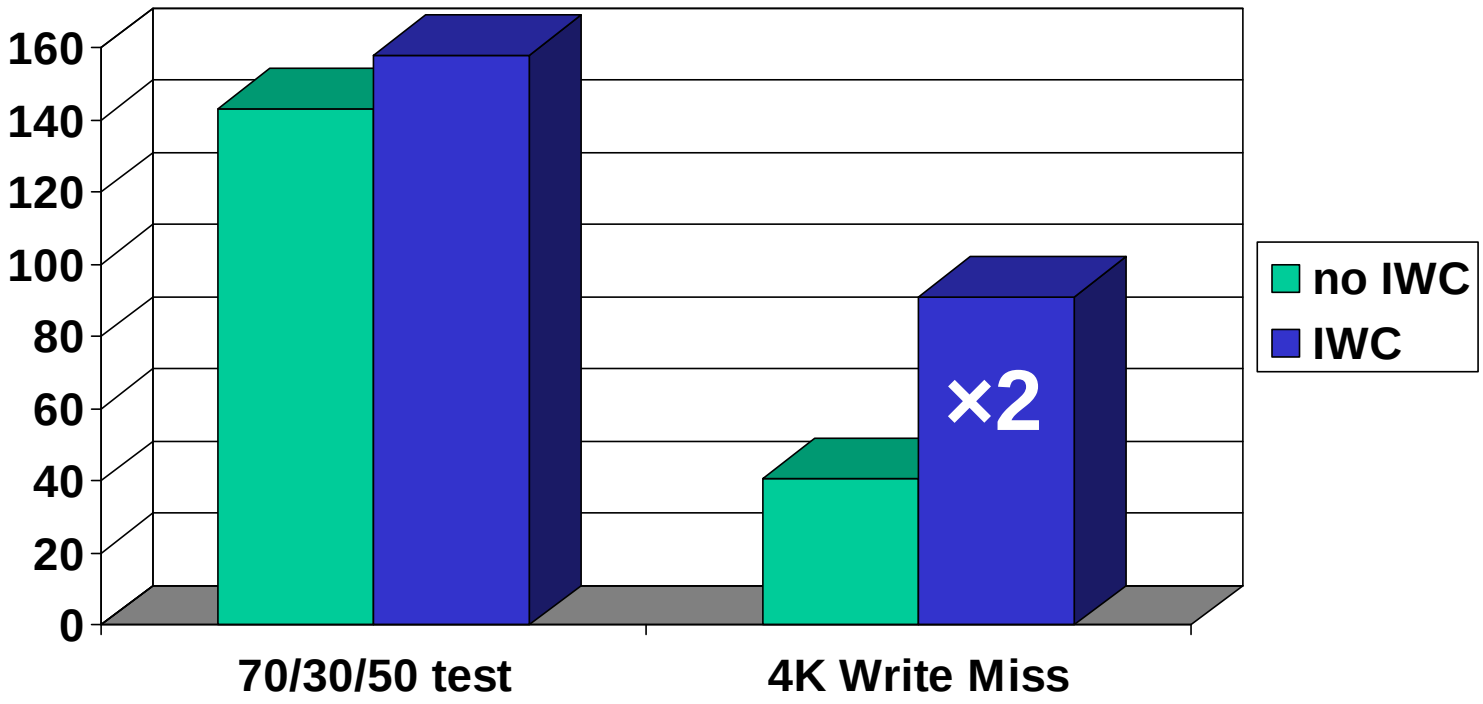
- Schreibaktivität optimiert für **minimale Kopfbewegungen**
- Ideal bei hohem Verteilungsgrad (große Pools, Striping)
- Verringerter Schreib-Overhead durch Ausnutzung der aktuellen Kopfposition

Intelligent Write Caching – Die Praxis



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.

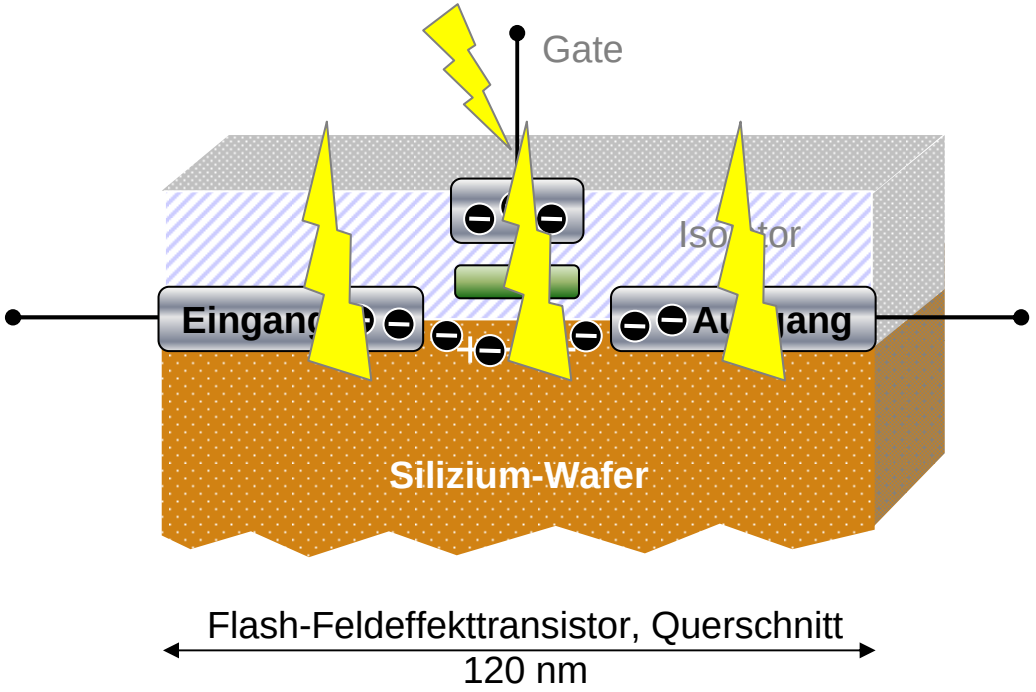
Intelligent Write Caching – für Datenbanken



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

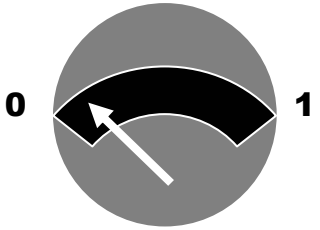
Disk versus(?) Flash

Funktionsweise von Flash-Speicher



Flash-Feldeffekttransistor, Querschnitt
120 nm

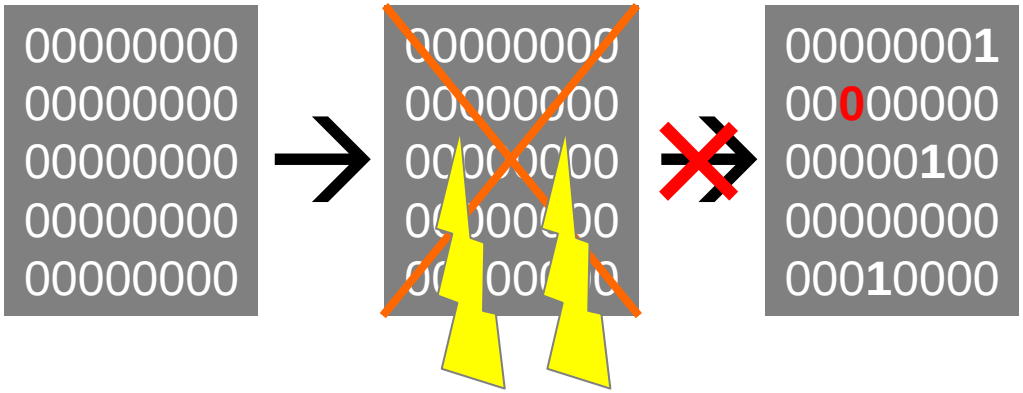
- **Floating Gate** (isoliert)



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

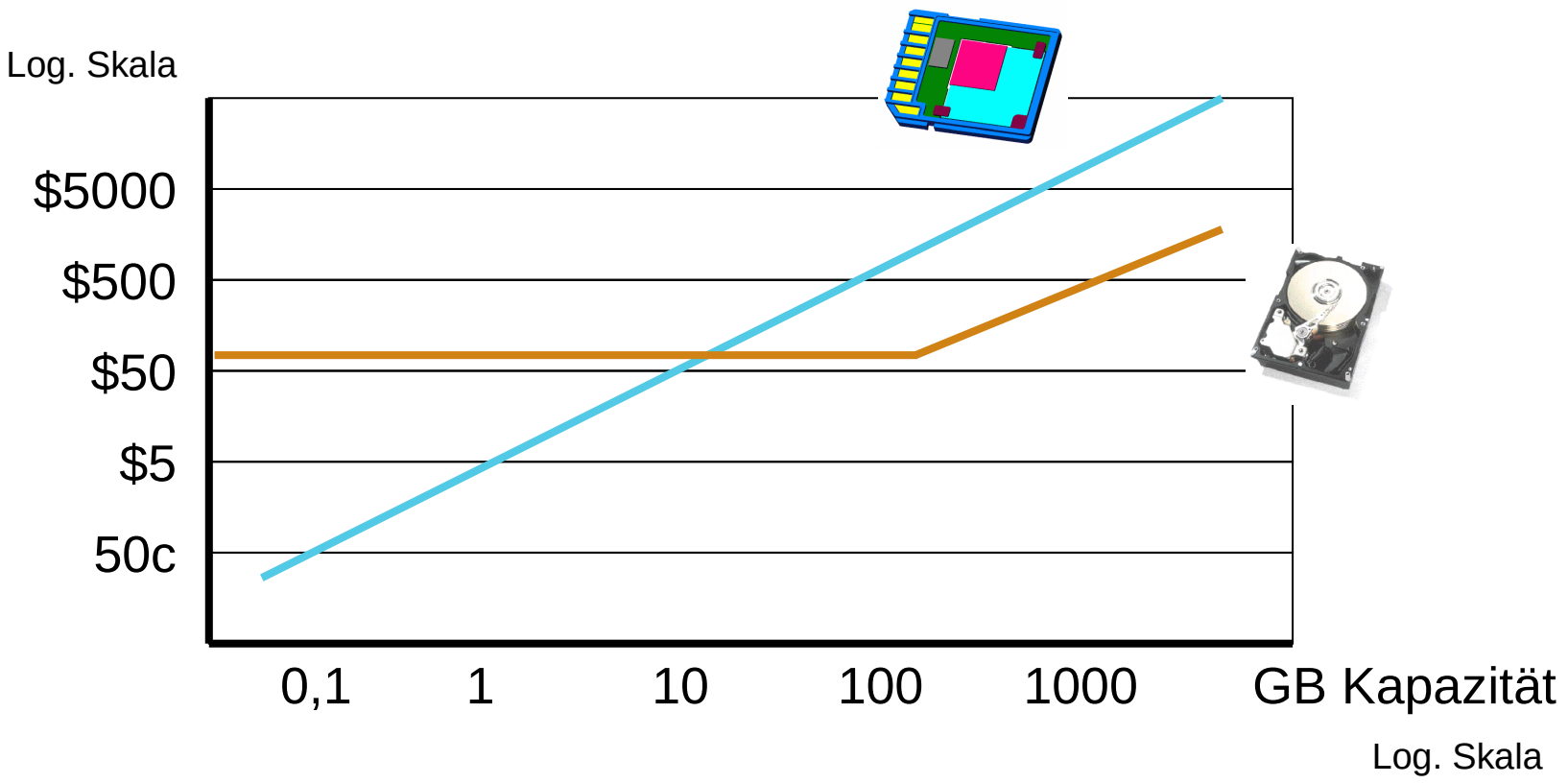
* Bei NAND Flash ist der aktive Zustand *NULL*, in diesem Fall sinngemäß "EINSEN" durch "NULLEN" ersetzten

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 Löschprozess **gesperrt**
- 10^5 Löschvorgänge je Block möglich, optimal streuen!

Preisentwicklung Flash versus Disk nach Kapazität



Nicht maßstabsgetreu; 2008 Schätzwerte

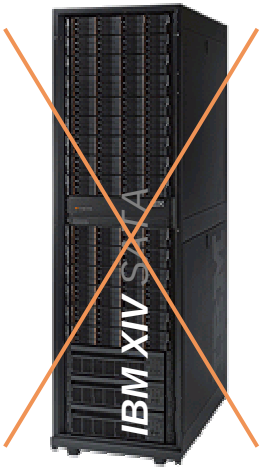
Solid State Disks SSD im IBM Portfolio

73/146 GB SSD
3,5" Carrier

64 GB SSD
internal

73/146 GB SSD
3,5" Carrier

Diverse Optionen
(Ankündigung demnächst)



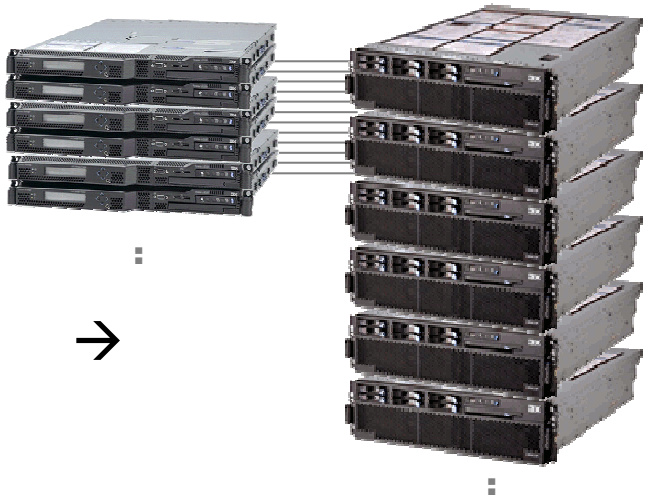
IBM SVC QUICKSILVER : 1 Mio *echte* IOPS

- 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 !

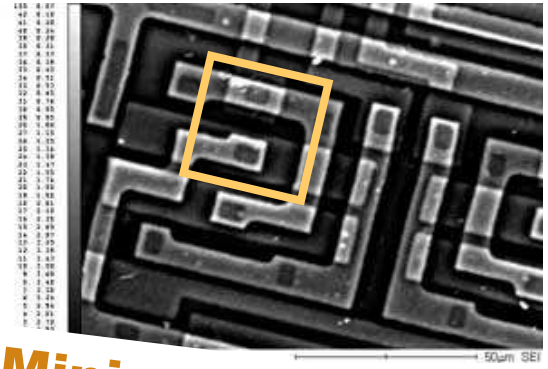
Alle Komponenten aus dem IBM Standardportfolio.

Flash von FusionIO®

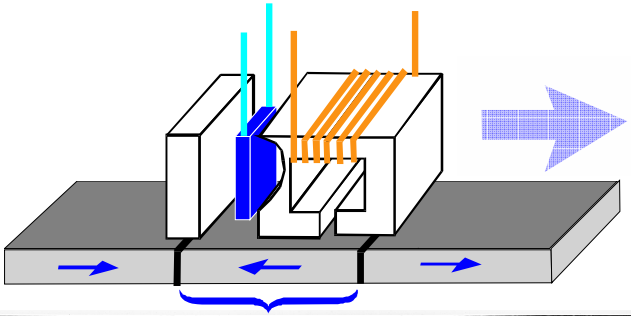
Chip- oder Analogspeicher?

Image © DOCB

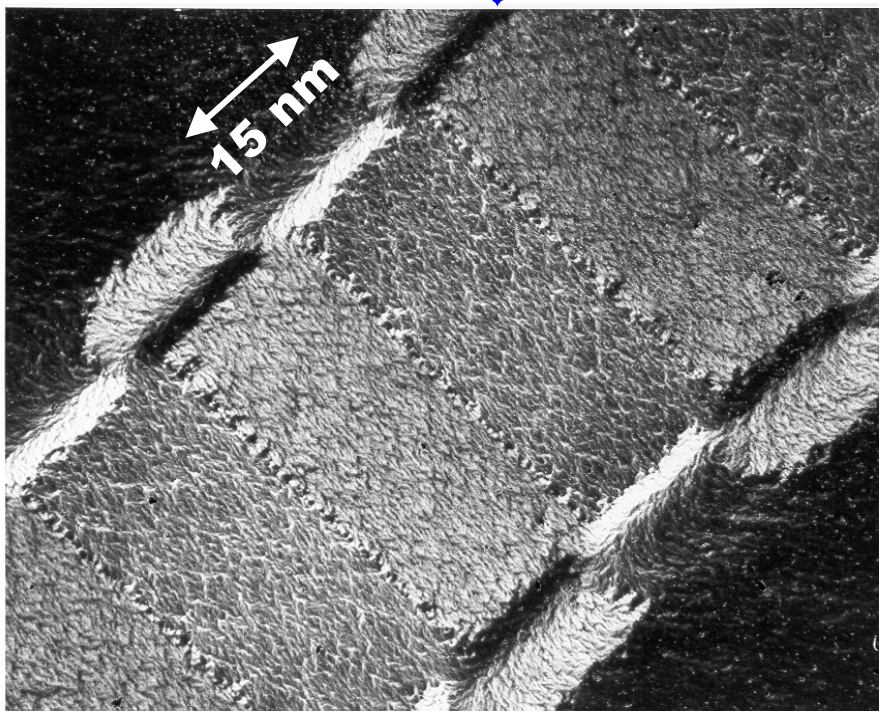
Chip-Speicher oder magnetischer Speicher ?



Minimum je Bit
ca. 40x40 nm

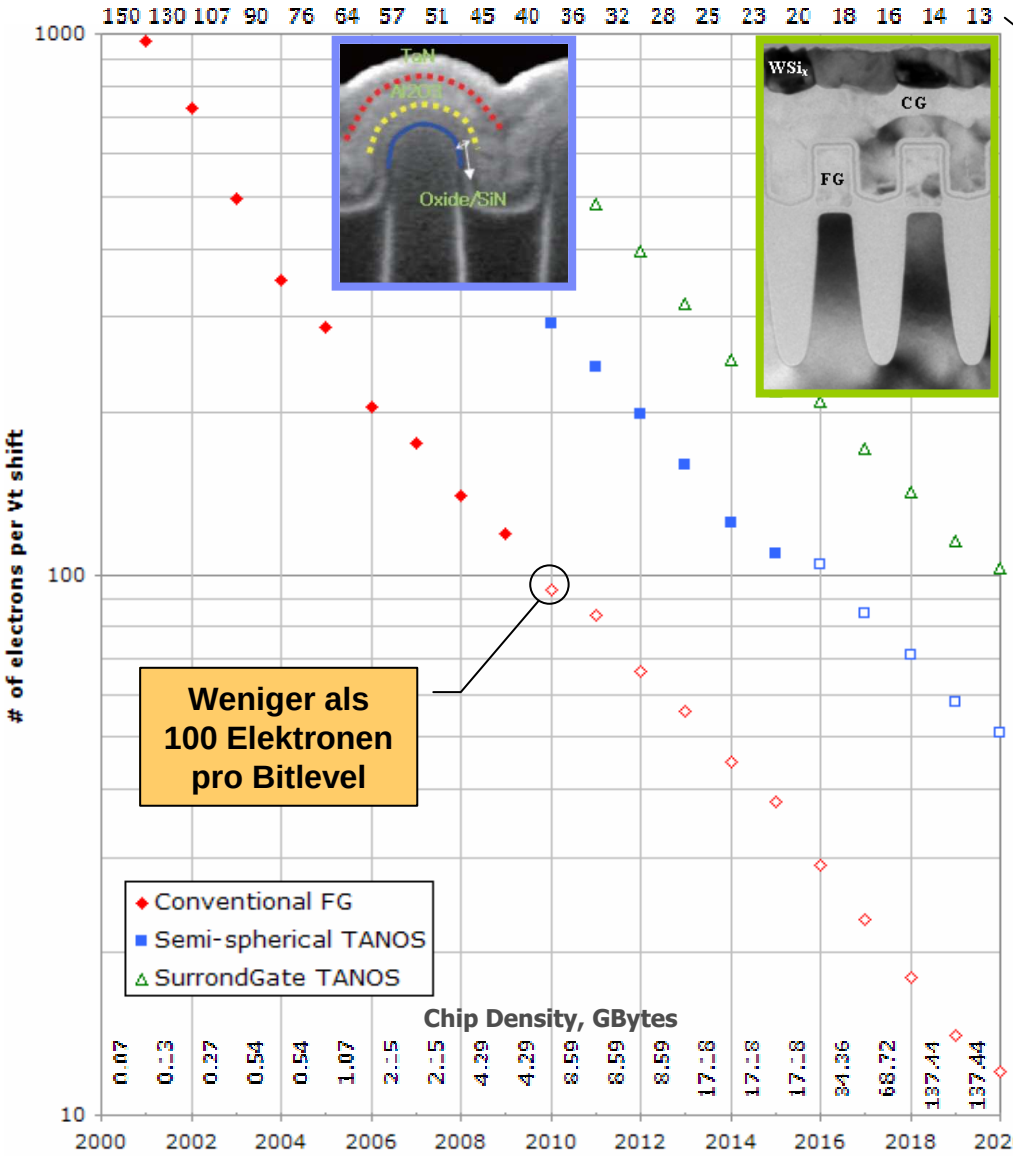


Disk :



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)*.

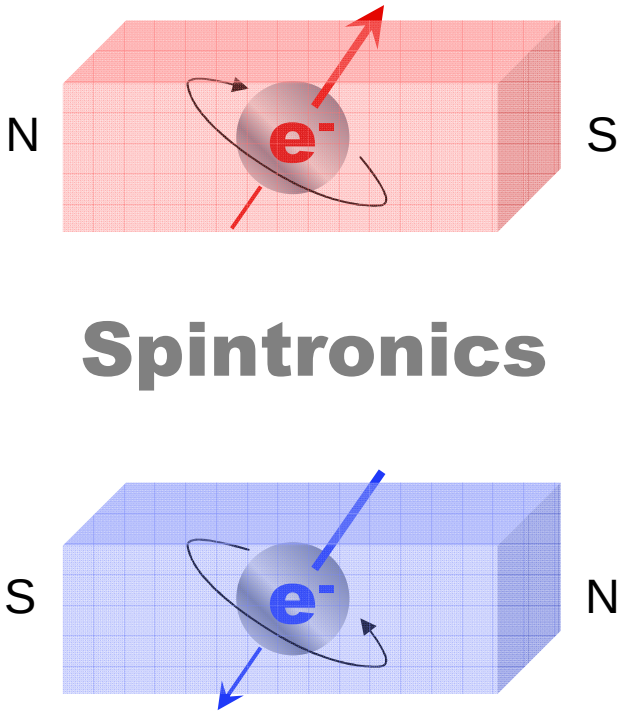
Pro Monat darf nicht einmal **ein Elektron** verloren gehen.

→ **Verschleiß-Anfälligkeit** steigt

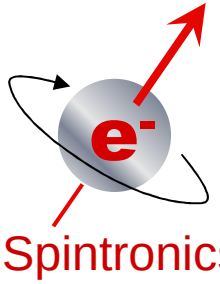
Ausweg: Mehr Elektronen trotz noch kleinerer Strukturen durch spherische oder zirkulare Gates (TaNOS = Tantal-Nitrid-Oxid-)

Quelle: IBM / Samsung

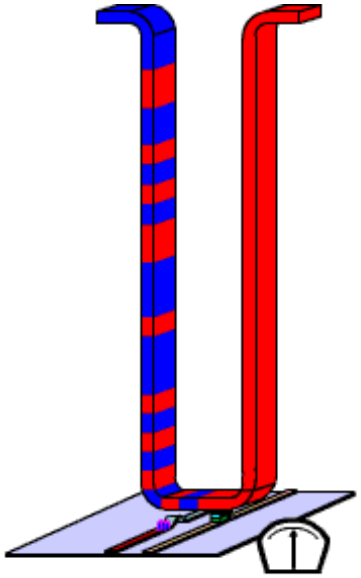
Spintronics: Neue Klasse von Festkörperspeichern

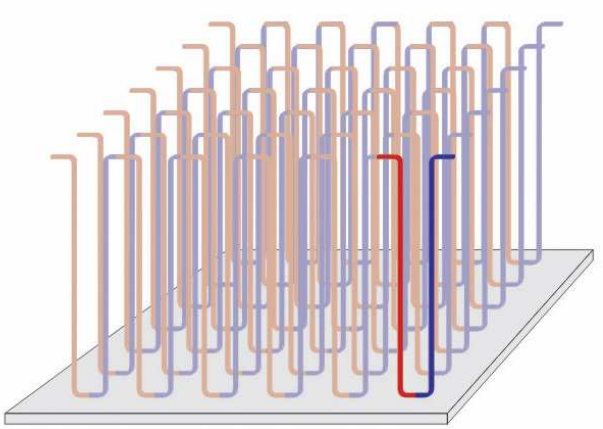


Speichern in 3D : "RaceTrack" für Spin-Elektronen

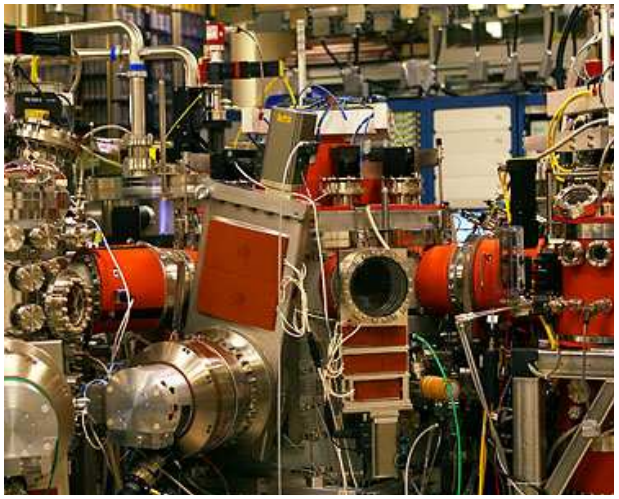


"Großer" Schreib/Lese-Kopf,
"kleine" Spin-Datenfelder auf
ferroelektrischem Nanodraht



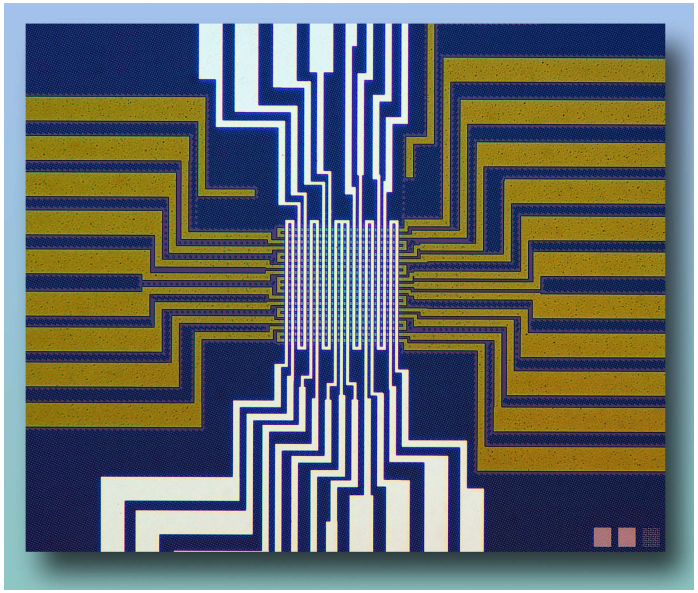


RaceTrack Storage Array:
Hohe Datendichte in 3D



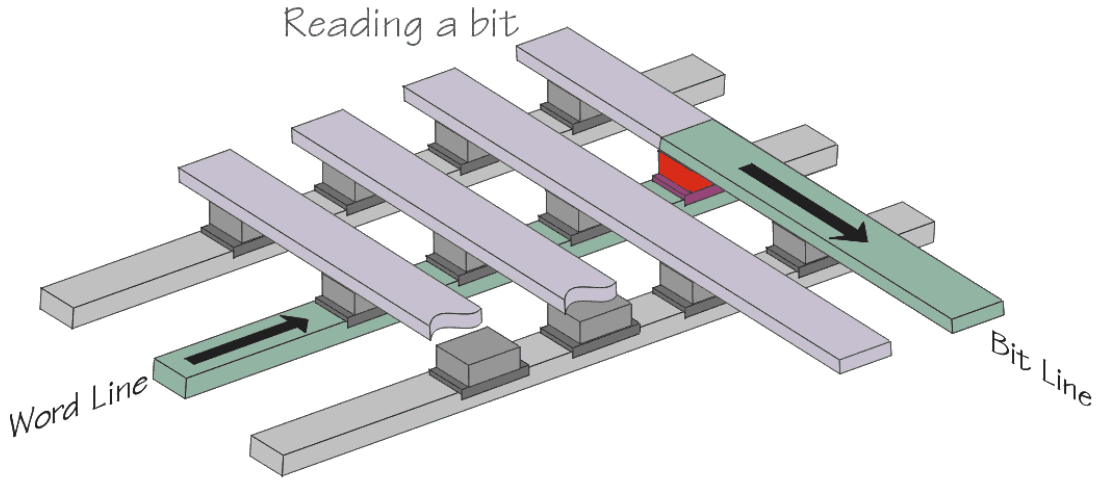
IBM Fellow Stuart Parkin,
Erfinder der GMR Leseköpfe,
erforscht "Racetrack Memory"

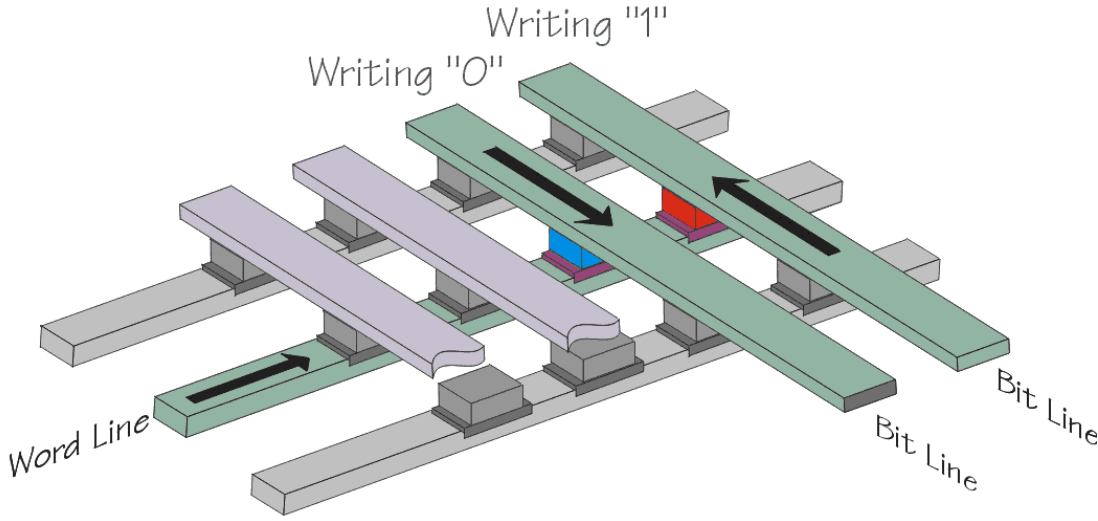
Magnetisches RAM



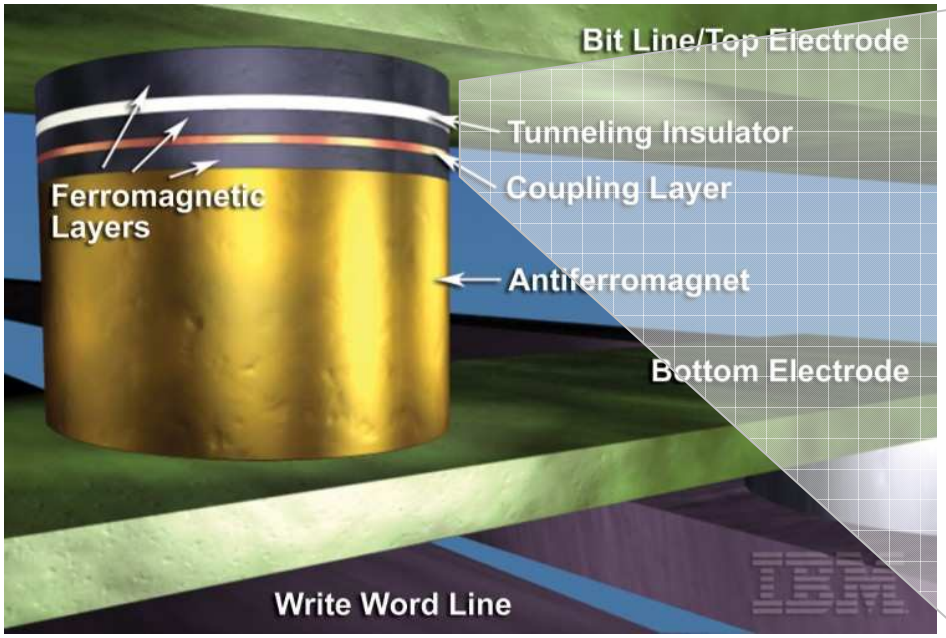
IBM Prototyp 199..

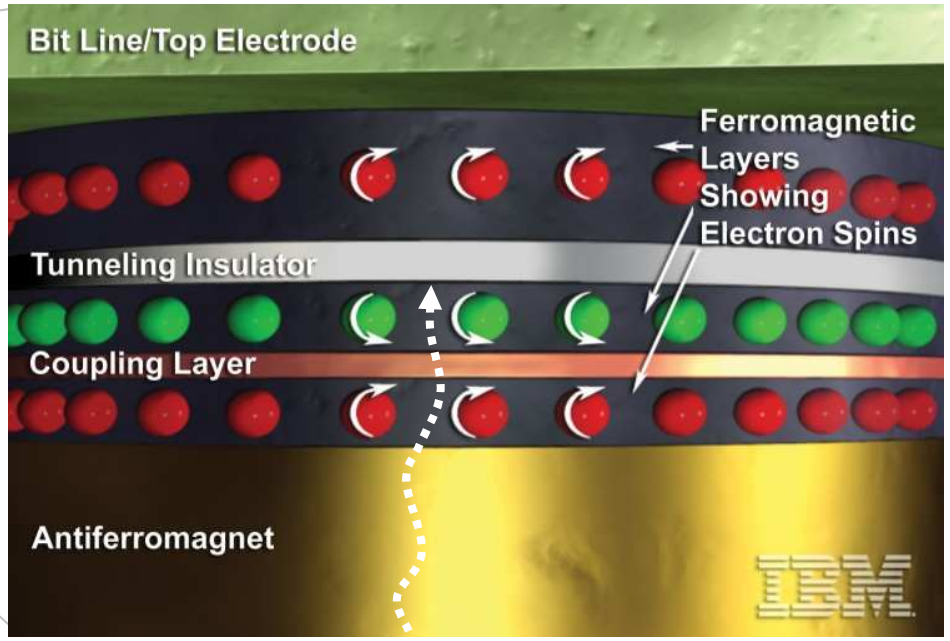
Bsp. Hersteller: Freescale Inc.
"MR2A16A" 4Mb nichtflüchtig
@ 35 nsec Zugriffszeit.

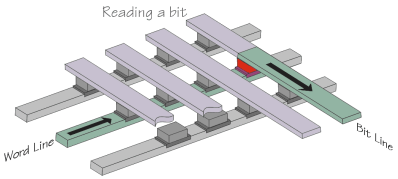




Magnetisches RAM



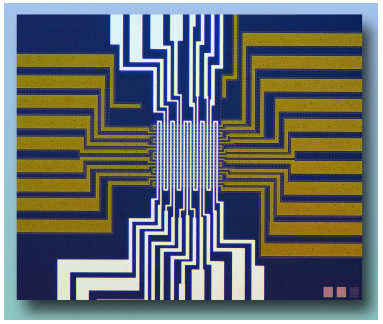


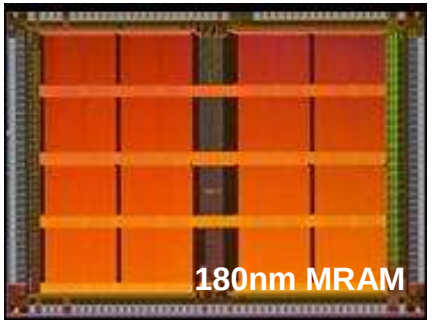


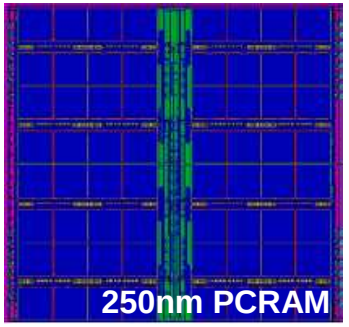
Neuer IBM Demonstrator:
2 nsec Zugriffszeit

(10 mal schneller als heutiges DRAM)

Nicht-flüchtiges RAM = IT Revolution !

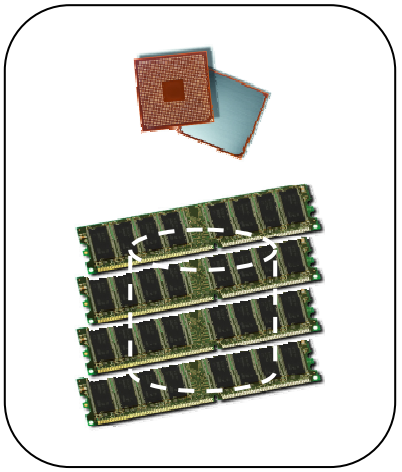




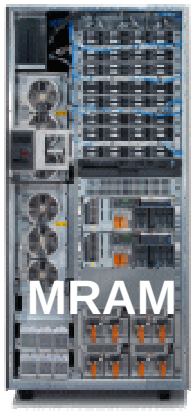


- Magnetic RAM / Ferroelectric RAM / Phase-Change RAM
- Flash Memory = Übergangstechnologie ("marktbereitend")

Wie verbaut man ***schnellen*** Solid-State Speicher?



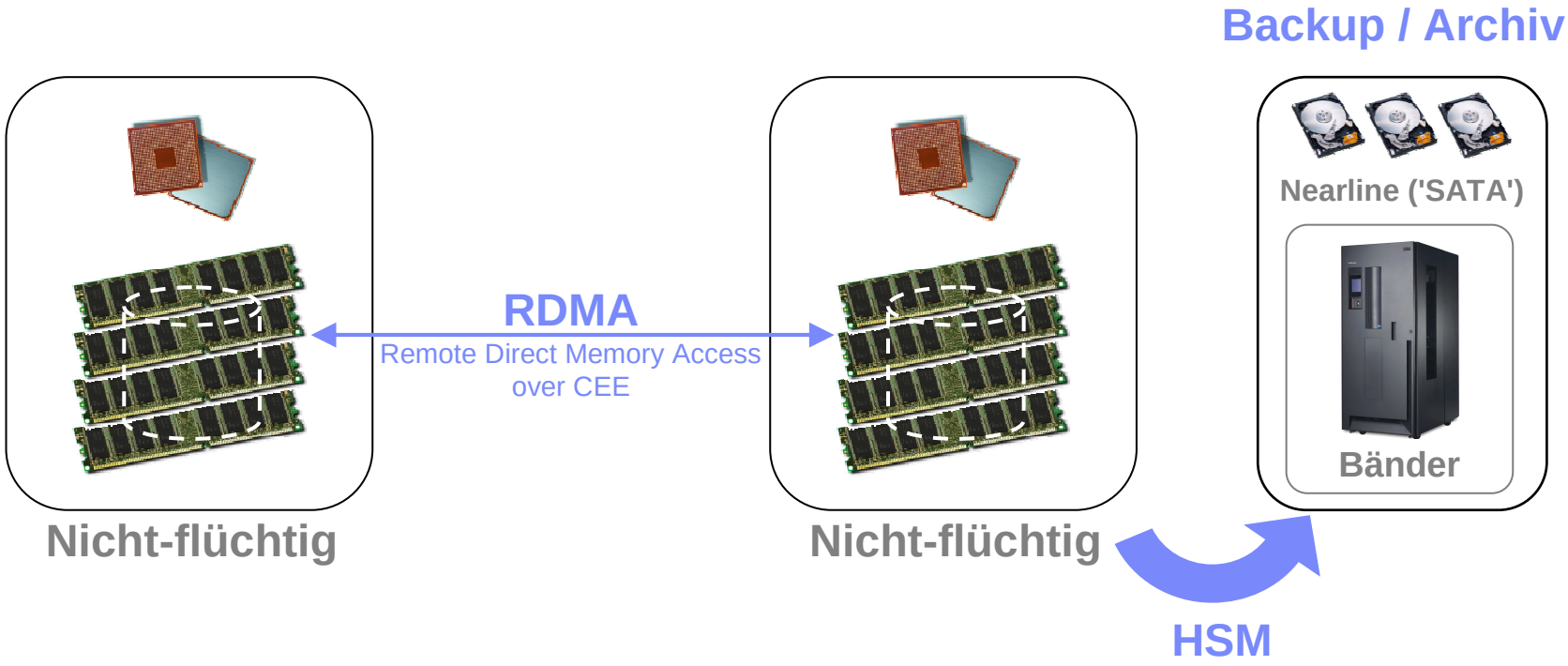
20 nsec



2 nsec

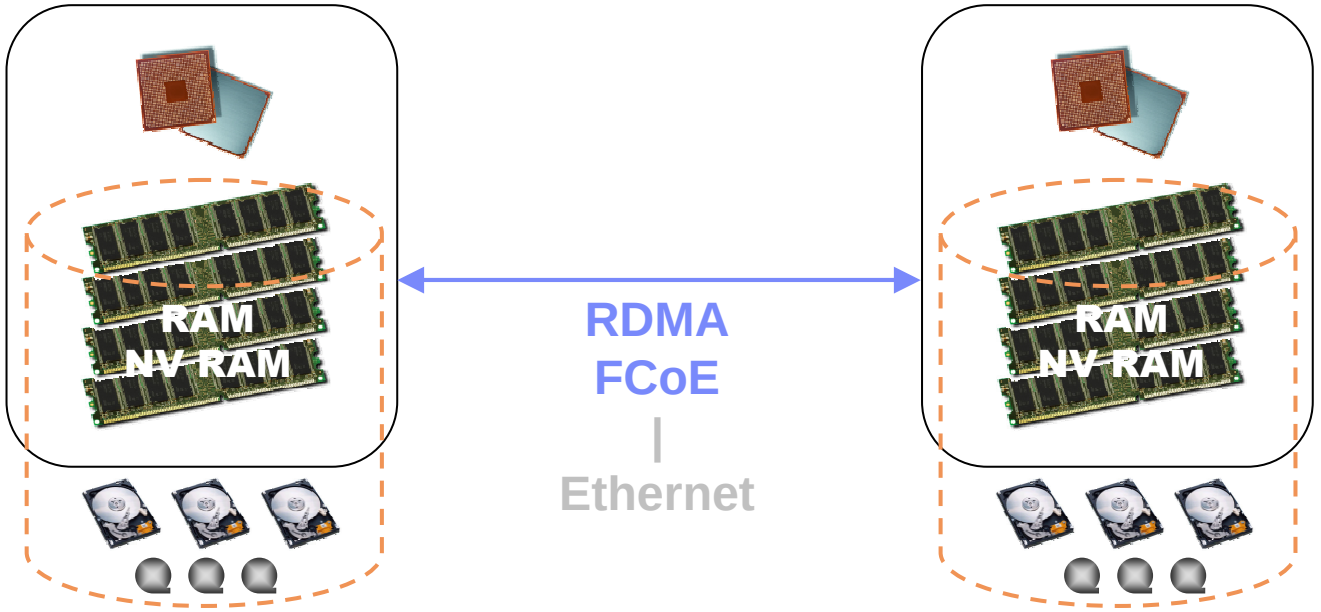
- Online Speicher zieht zum Prozessor
- Schnelle (Fibrechannel-) Platten verschwinden
- Monolithisches Design = hohe Verfügbarkeit

Die SANs der Zukunft



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

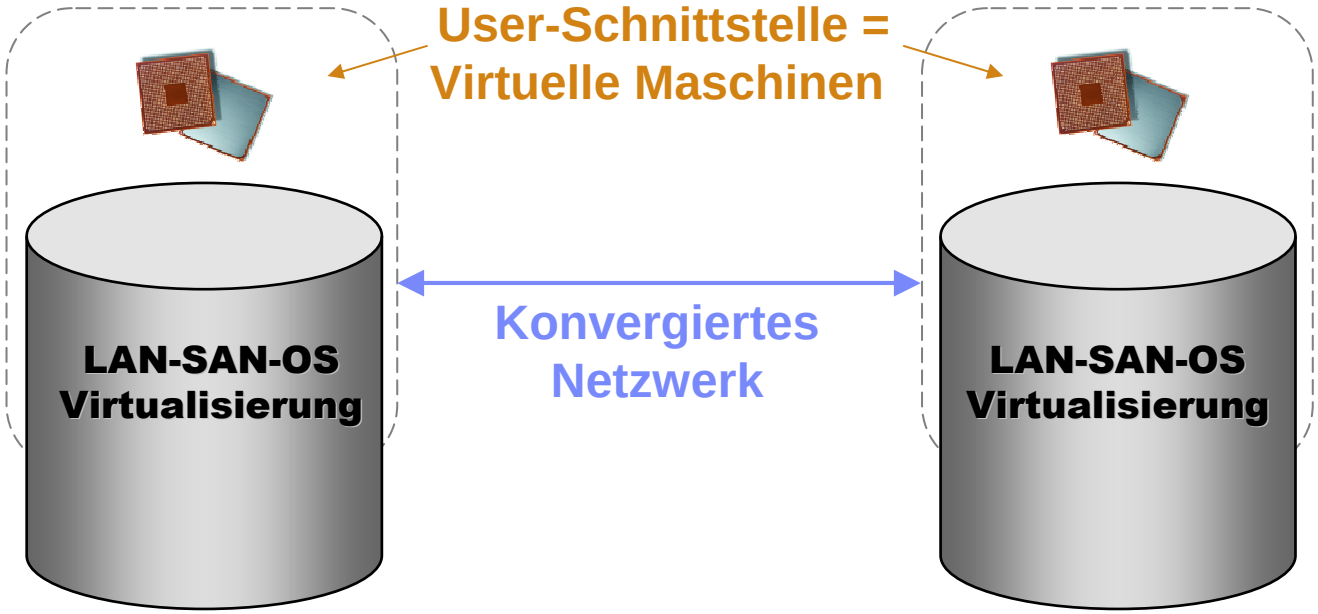
Storage in der Zukunft : *Multi-Tier* Speicher



- Das konvergierte Netzwerk verknüpft *alle* Speicherklassen von RAM bis Tape, nebst Nachrichten- und Dateiübertragung.

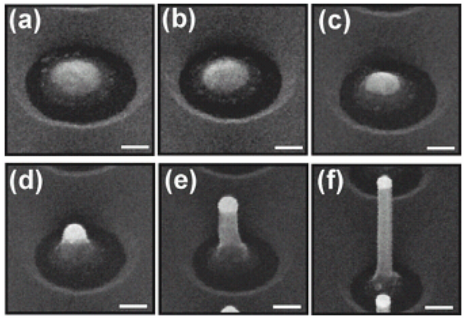
- Applikationen sehen nur noch RAM und RAM LUNs (zwecks Kompatibilität). Hierarchieverwaltung verläuft davon getrennt.

VM Speicher/Server

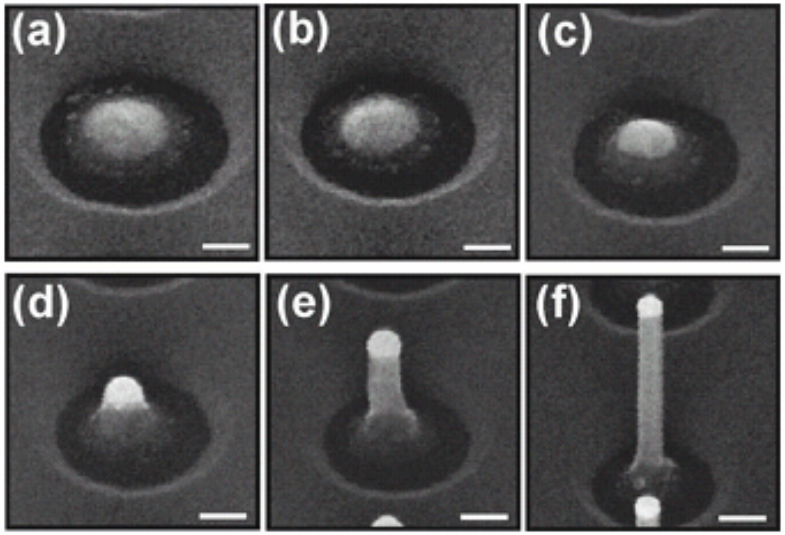


- Regelbasiert
- Komplexität wird vom Hersteller gehandhabt
- Software hat keine Systemsicht mehr

Die Nano-Zukunft

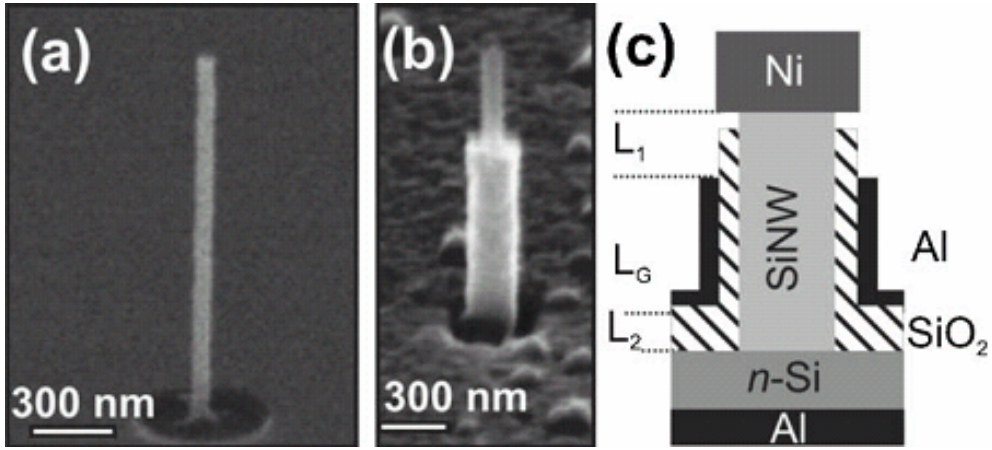


Die schnellsten & sparsamsten Transistoren



← Wachstum von Silizium-Nanosäulen

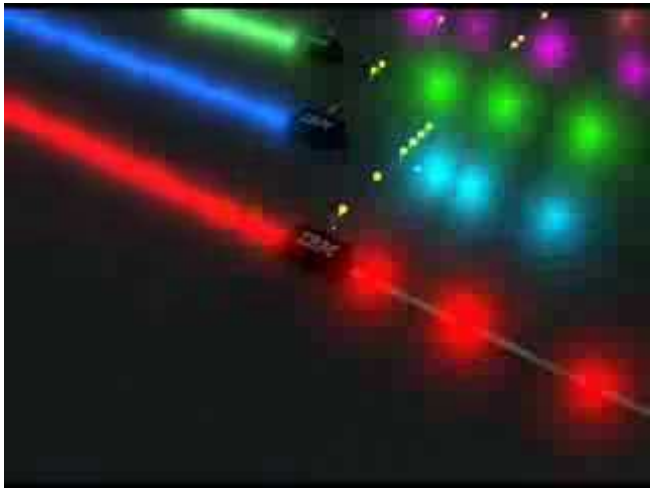
Anwendung: Schnell sperrender Transistor



Mit Mantel-Gate versehene Nanosäule

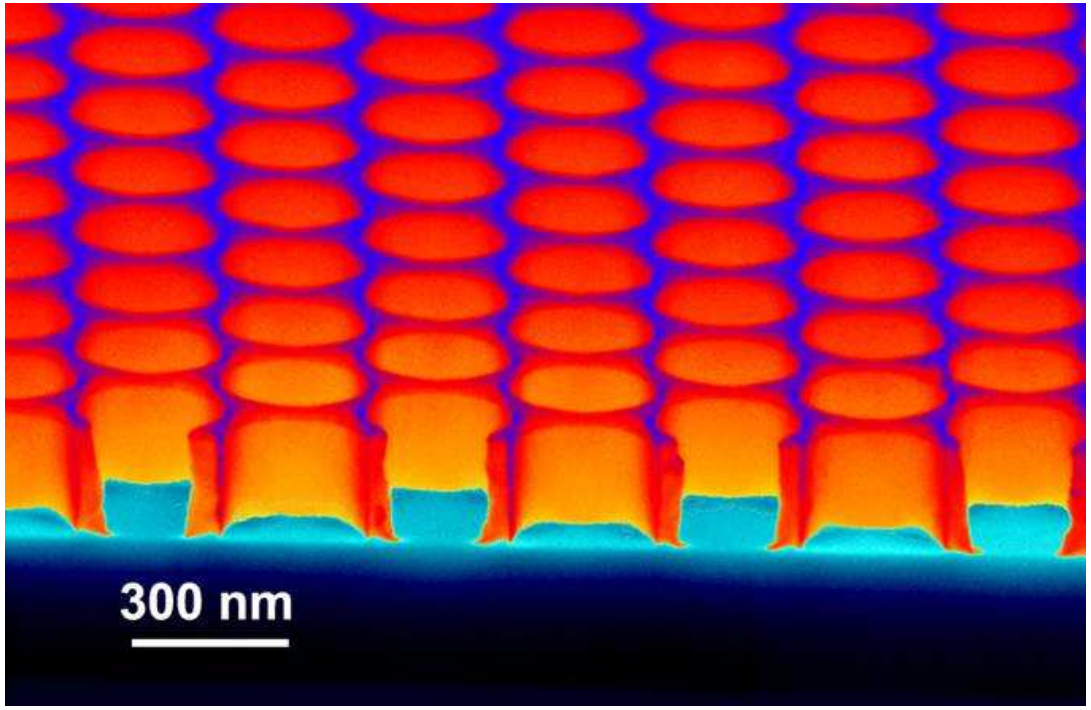
- Minimaler Verbrauch
- Höchste Schaltgeschwindigkeit
- TeraHertz nicht ausgeschlossen

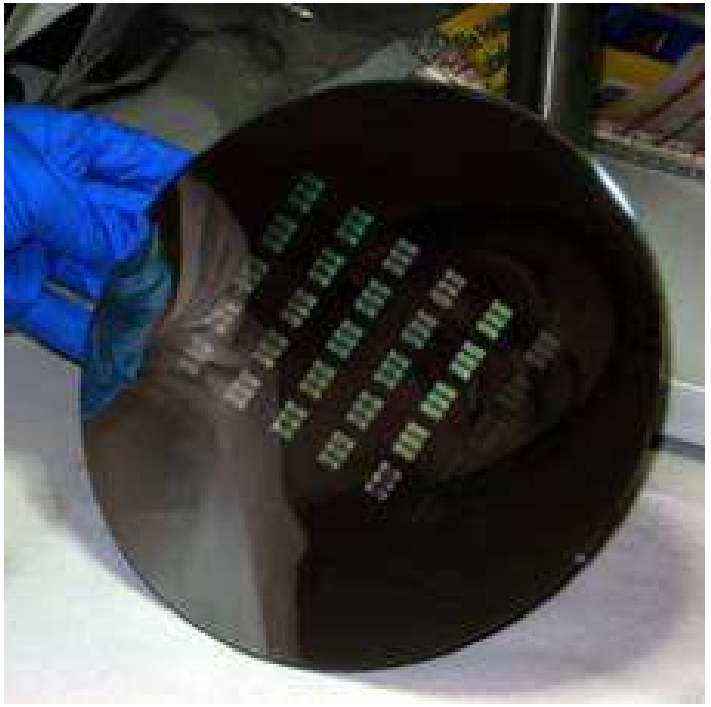
Die optische Zukunft



Langsames Licht

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





axel.koester@de.ibm.com

Disclaimer