IBM Systems Storage

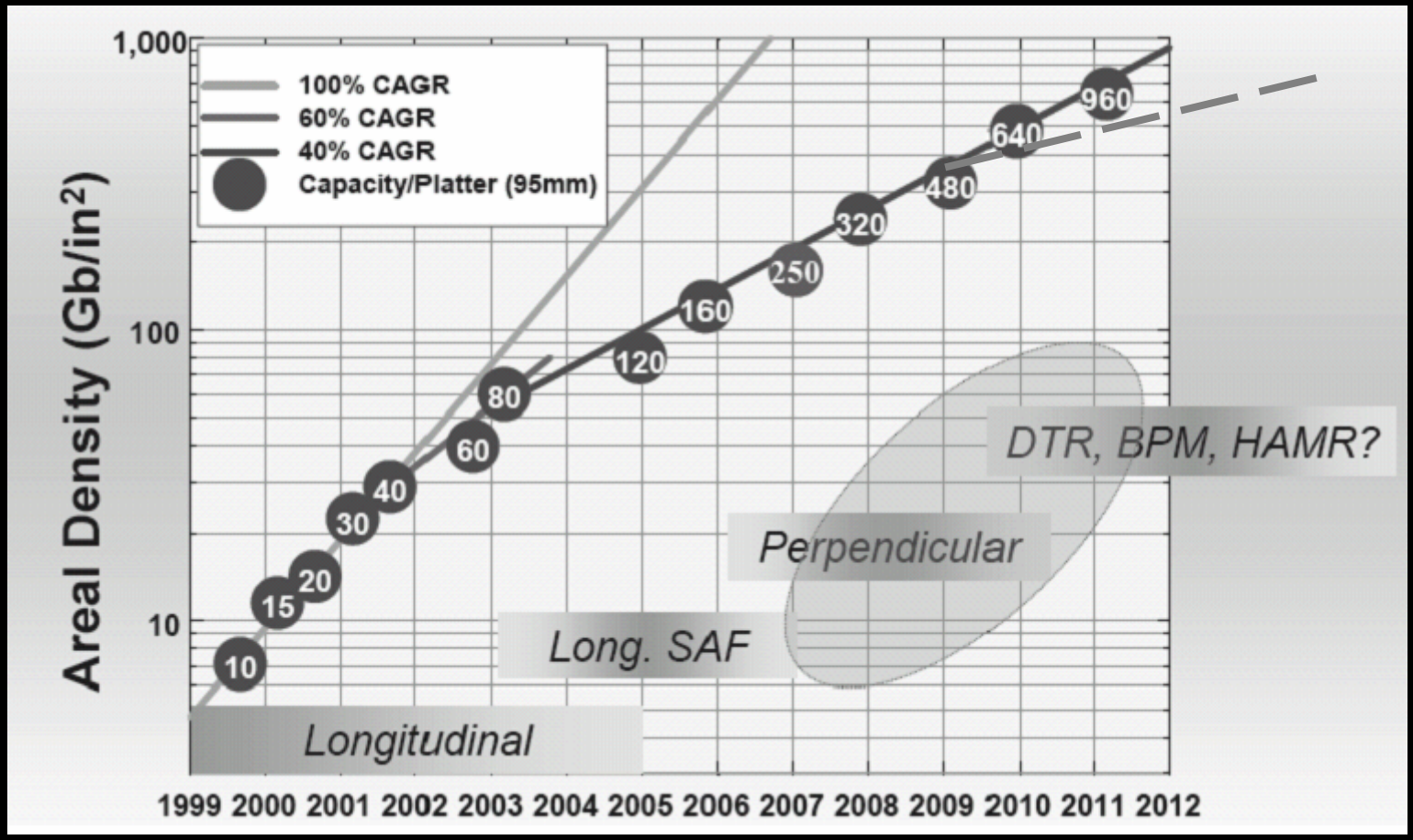
Storage Trends
Technologieausblick

Dr. Axel Köster
Senior Consultant
IBM Enterprise Storage
axel.koester@de.ibm.com

Themen

1. Ausblick FC-, SAS-, SATA- Festplatten
2. Flash Technologie
3. Auswege aus dem Festplattendilemma
4. Neue Speichertechnologien

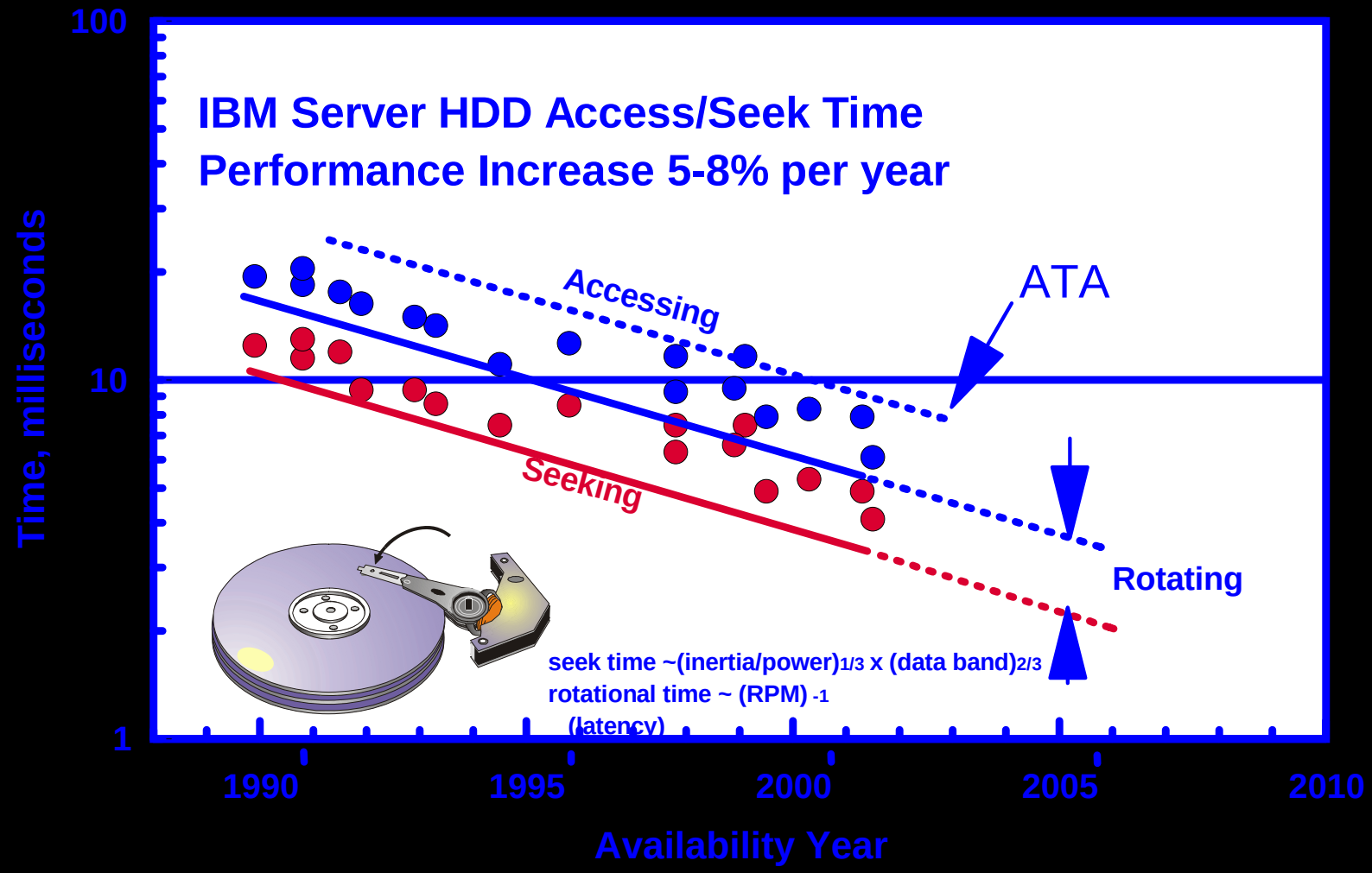
Aufzeichnungsdichte-Steigerung +30…40% pro Jahr



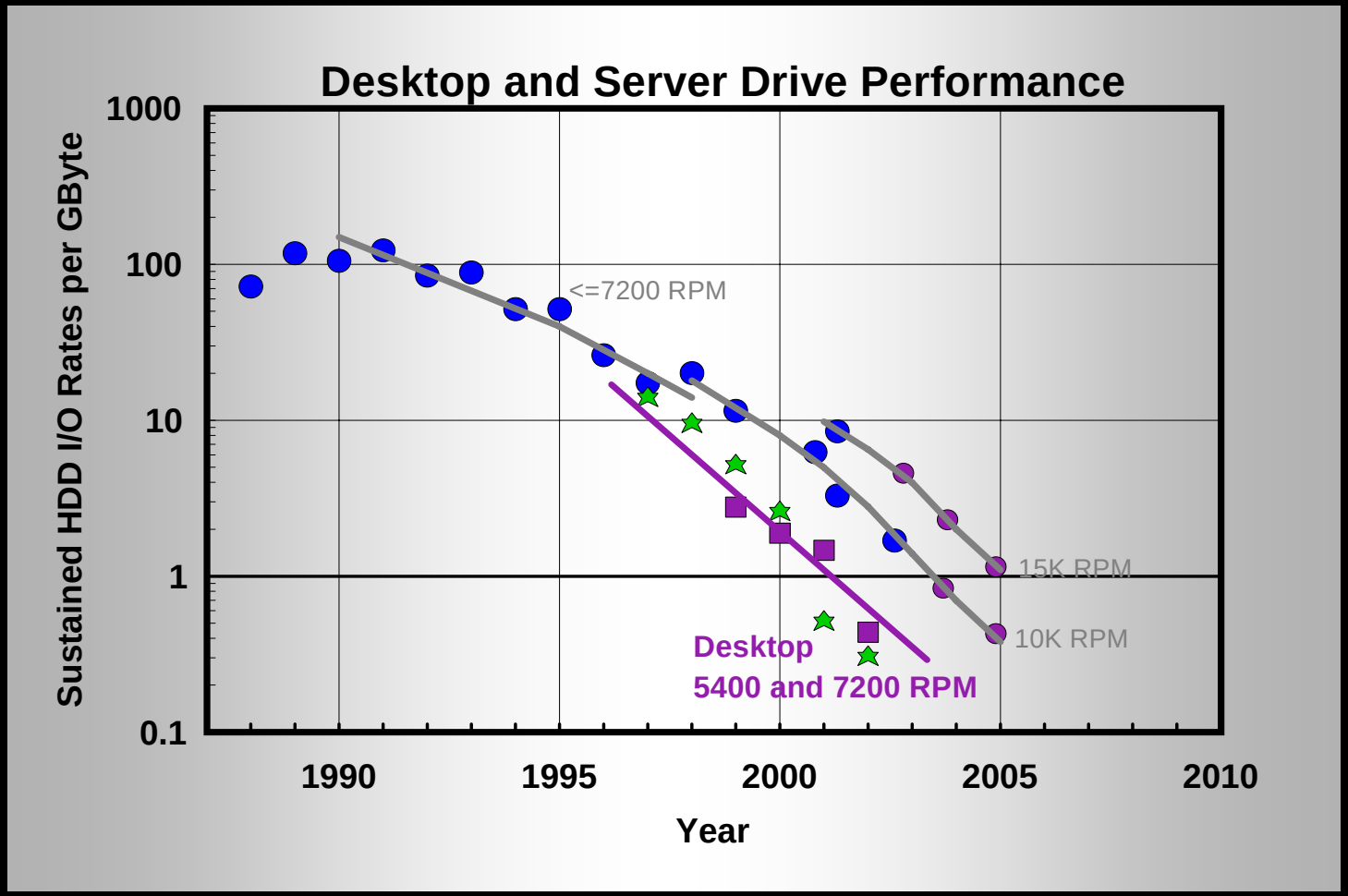
Source: Komag, IDEMA Symposium

Bit Patterned Media (BPM), Heat Activated Media Recording (HAMR)
Seagate glaubt an 2670 GB per 3.5” disk und 100 GB per 1.0” disk bis 2013

Problem: Zugriffszeit von Festplatten hält nicht Schritt



Datenrate *pro Gigabyte* fällt dramatisch

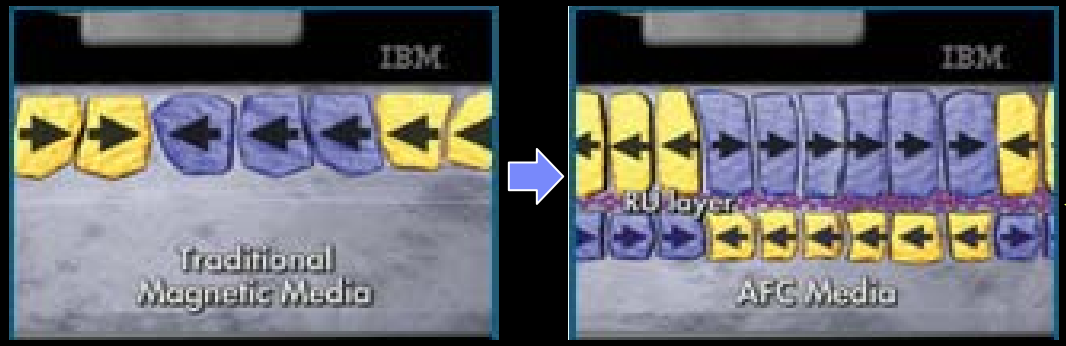


Schnellerer Zugriff? Nicht im 3,5" Zoll Format

- 3,5" Platten mit >15.000 RPM sind nicht preiswert herstellbar. *(Luftreibung* = $\frac{1}{2} \rho \cdot c_w \cdot A \cdot v^2$*)*
- 2,5" Platten werden zum Enterprise-Standard, evtl. noch 1,8" mit 20.000 RPM (fraglich!)

1" Microdrive mit 20.000 RPM?

Physikalische Dichtegrenze der Magnetisierung

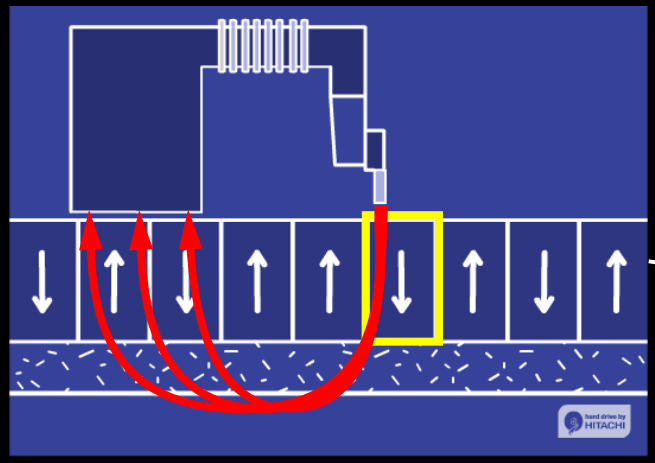


>100.000 Spuren pro Zoll

Antiferromagnetische Kopplung durch 3-Atomlagen Ruthenium

"Pixiedust" stabilisiert kleinste Bits, die sonst durch Umgebungswärme ihre Ausrichtung verlieren

"Perpendicular" Aufzeichnung auf klassischen Medien



Pixiedust AFC Drives

15kRPM-taugliches Medium

Fibrechannel, SAS

Perpendicular Drives

Aufzeichnung bis 7200 RPM

SATA, F-ATA

- Schreibe up/down Bits mit großer Eindringtiefe
- Großes Bit unter kleiner Schreibzone → längere Verweildauer nötig
- CPP-GMR Kopftechnik mit 500 GBit/inch² Demonstrator (3.5" ~ 4TB)

Source: Hitachi GST (former IBM San Jose Storage Systems) | CPP = current perpendicular to plane | GMR = Giant magneto-resistive

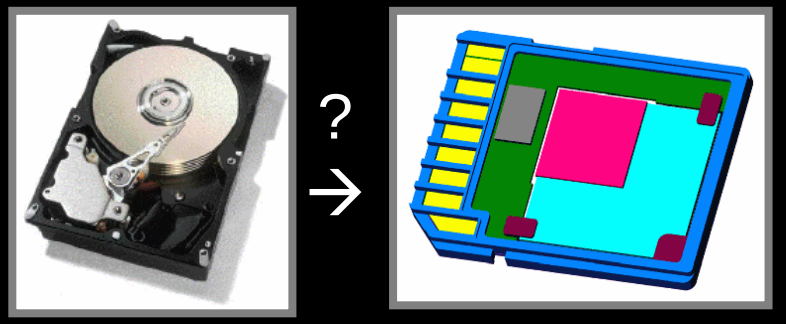
Themen

1. Ausblick FC-, SAS-, SATA- Festplatten
2. Flash Technologie
3. Auswege aus dem Festplattendilemma
4. Neue Speichertechnologien

Wann ersetzt Flash Speicher die Platten?

Herausforderungen:

- Schreibgeschwindigkeit
- Langlebigkeit der Zellen
- Günstigerer Preis



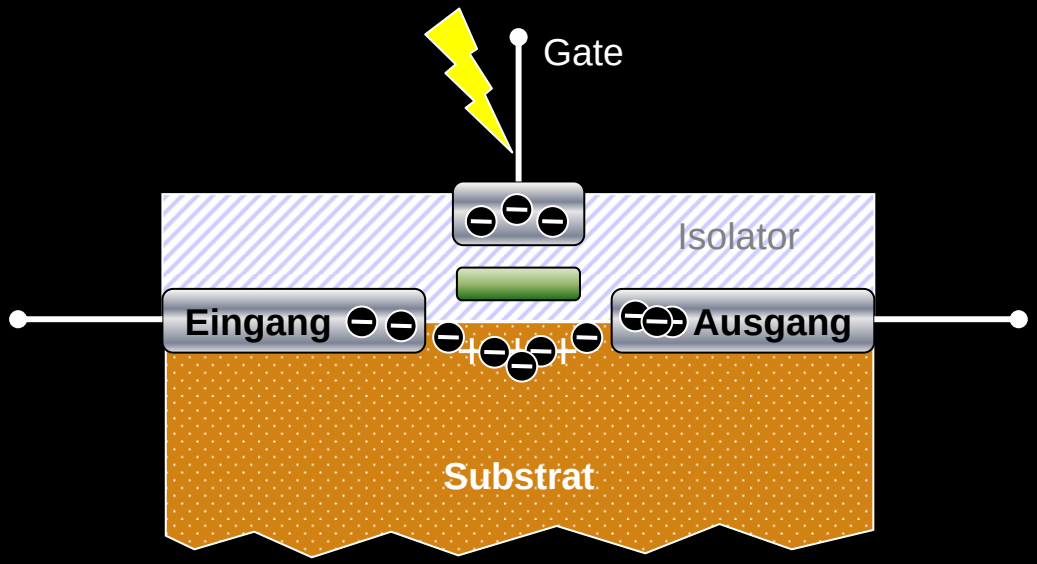
Samsung™ SSD geöffnet
(Altes Modell 4GB)

IBM 16/32GB SSD Laufwerk
für Bladecenter & System x
(2,5" SATA Interface)

Preislage $30 *pro Gigabyte*

ADTRON™ SSD 3,5" + 2,5" 160GB
Preislage $80-$115 *pro Gigabyte*

Funktionsweise von Flash-Speicher *(EEPROM)*

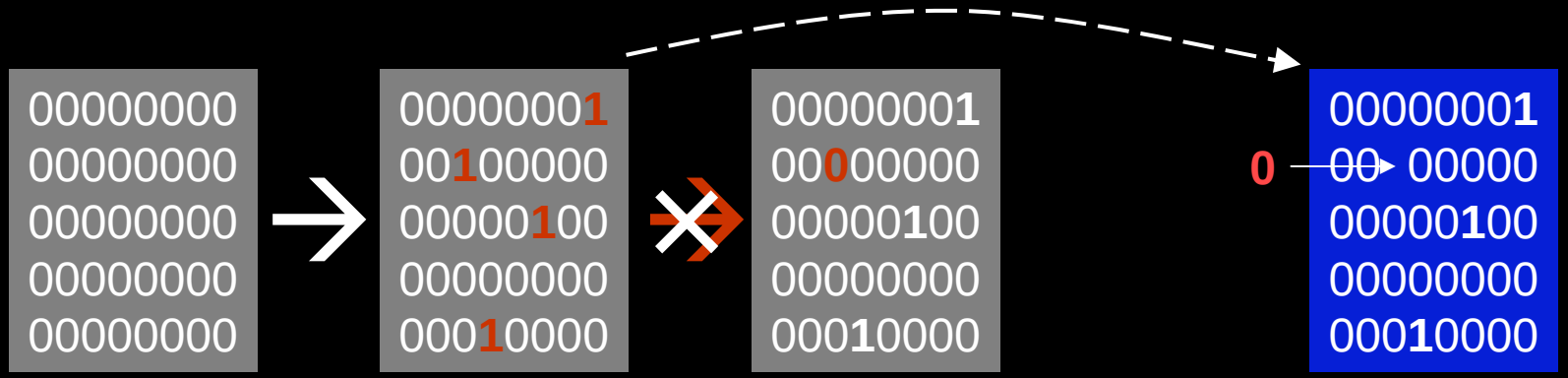


Querschnitt durch Flash-Feldeffekttransistor

- Es können nur Einsen geschrieben werden
- Individuelles Löschen nicht möglich, nur Blocklöschung
- Beim Löschen altert das Substrat (max. 1 Mio mal)

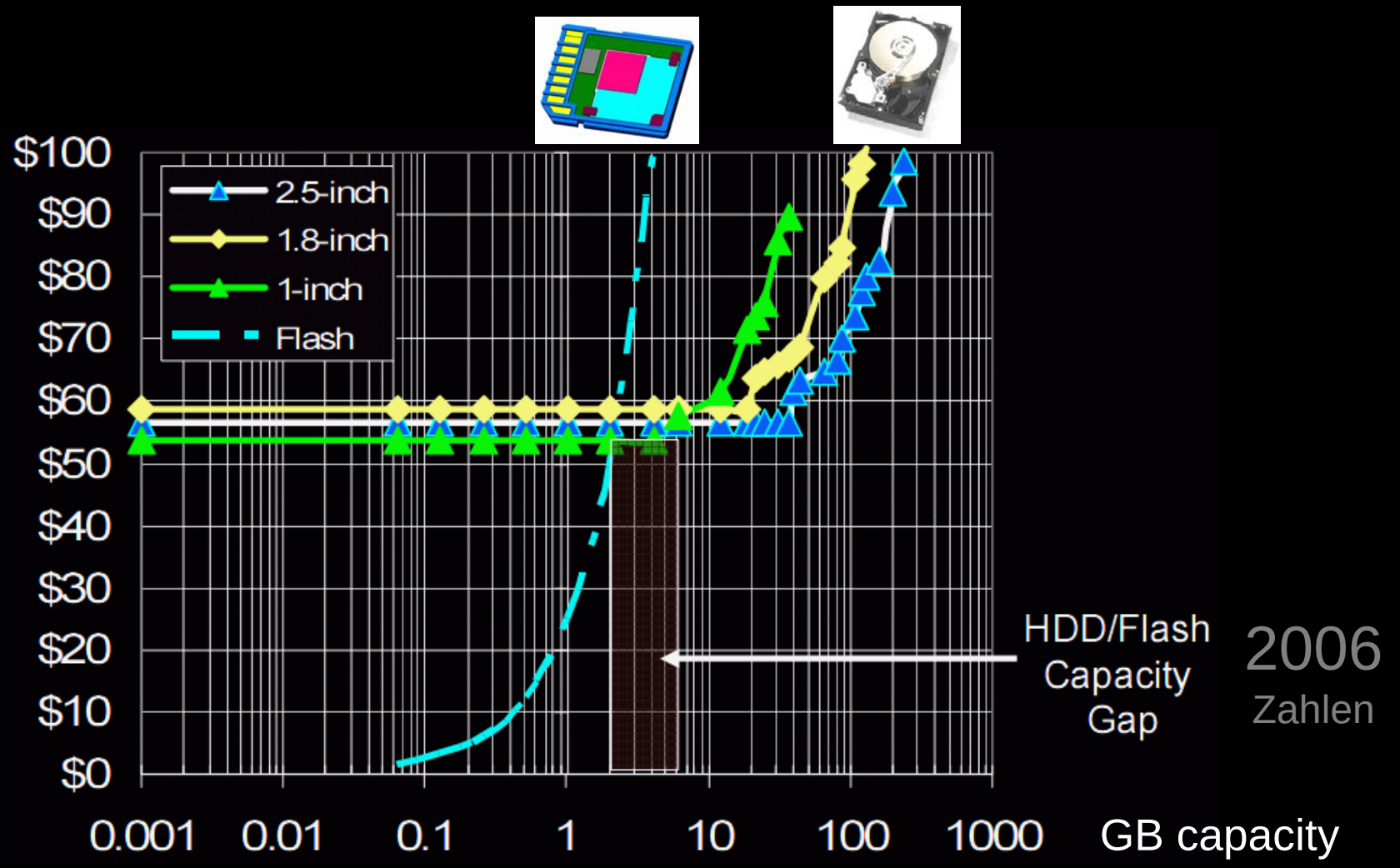
Multilevel-Flash
(4…16 Bit pro Zelle)

Beschreiben & Löschen von Flash-Speicherblöcken



- *Random Write* ist nicht der optimale Workload für Flash
- ***Random Re-Write*** im gleichen Block ist zu vermeiden !
  - ➔ DRAM-Cache vorschalten !
- Das Verlagern oft genutzter Blöcke verringert vorschnelle Alterung

Preisentwicklung Flash versus Disk nach Kapazität



Source: Coughlin Associates 2006

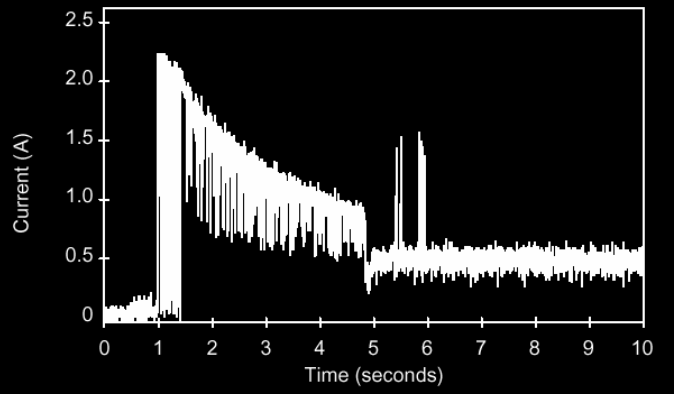
Hybrid Laufwerke

HDD mit integriertem Flash Bereich (HHD)

- Samsung 80-160GB, 128/256MB Flash
- Seagate 160GB 5kRPM, 256MB Flash
- Hitachi…
- Fujitsu…

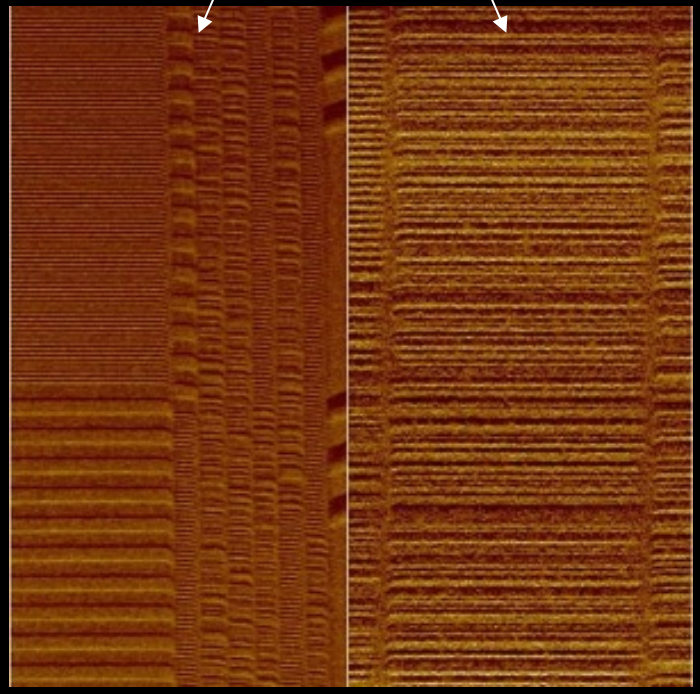
- Vorteil:
  - Hybrid-HDDs laufen seltener an, da sie Einzelschreibvorgänge im Puffer auffangen
- Herausforderungen:
  - Vergleichbare Zuverlässigkeit
  - Preis (derzeit +20…30%)

HDD Anlauf-Stromaufnahme



A propos Preisentwicklung: Tape ist konkurrenzlos

Magnetische Kraftmikroskopbilder von
8 TB Band vs. LTO-3



6,67 Gbit pro square inch ~ 1,033 Gbit pro cm^2

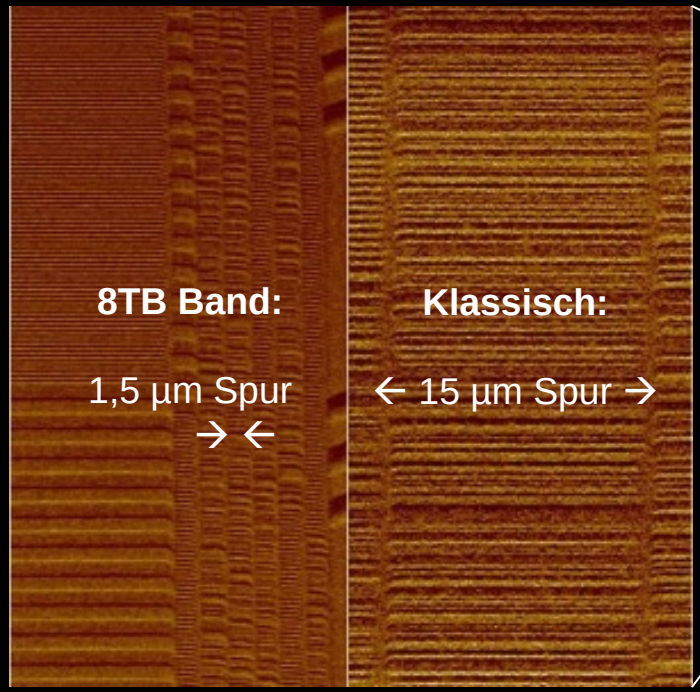
15-fache Speicher-dichte heutiger LTO3

- Demonstration bei Produkt-typischer Bandgeschwindigkeit (4 Meter/sec)

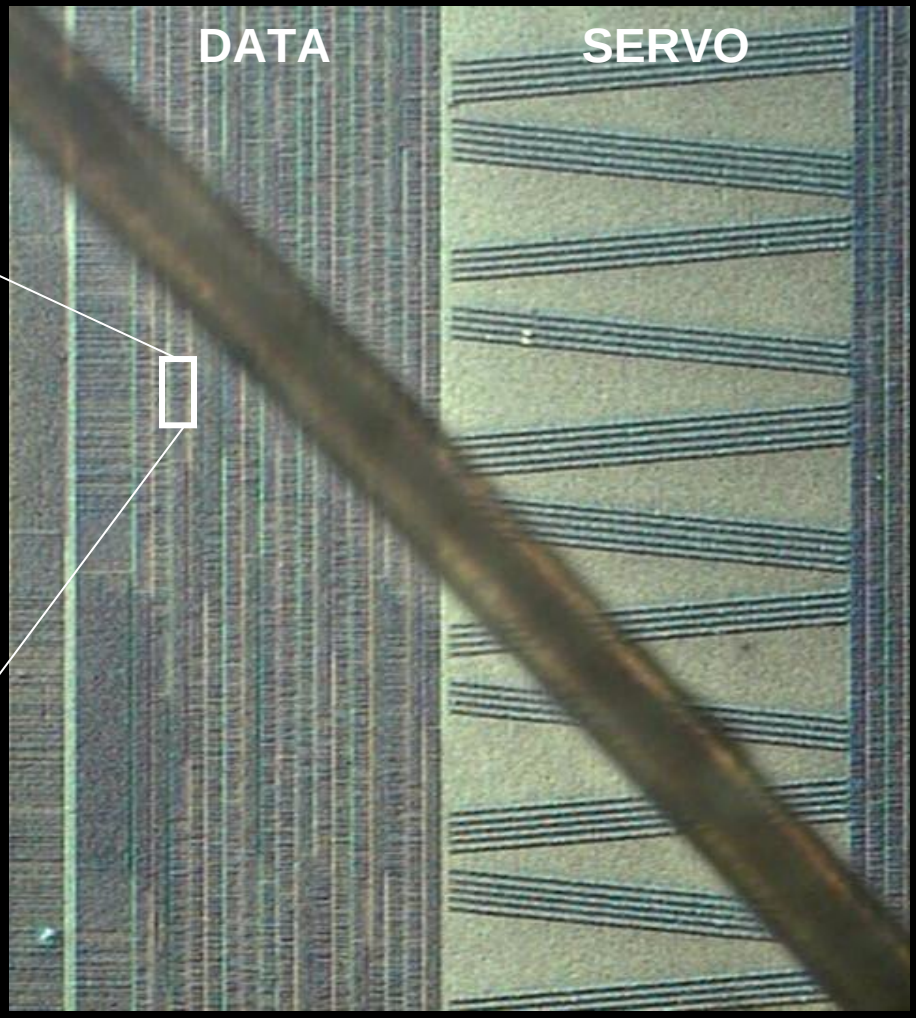
> 1 Gigabit/cm^2

8 TB auf Standard ½" Cartridge

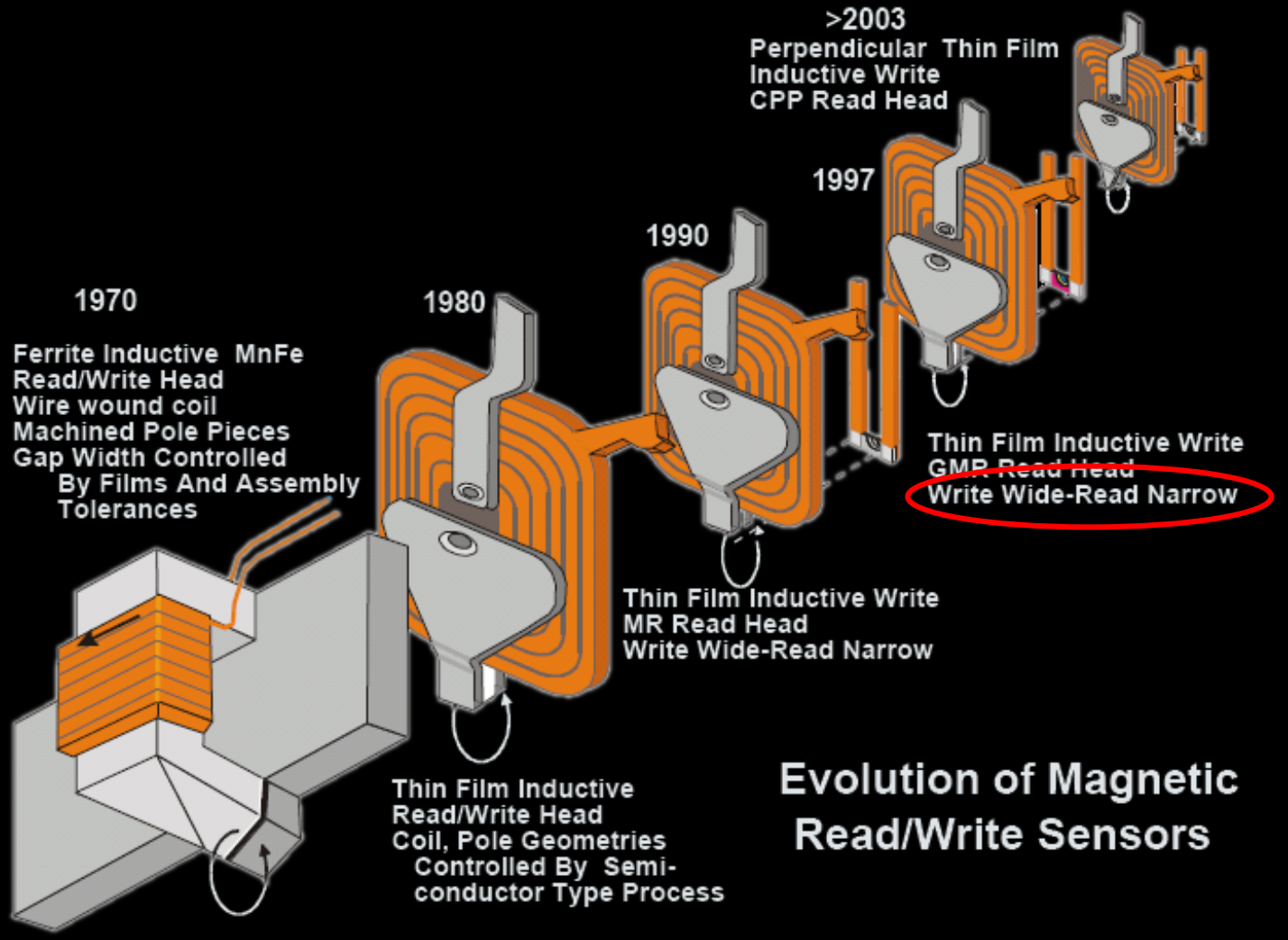
IBM TS1120 Jaguar



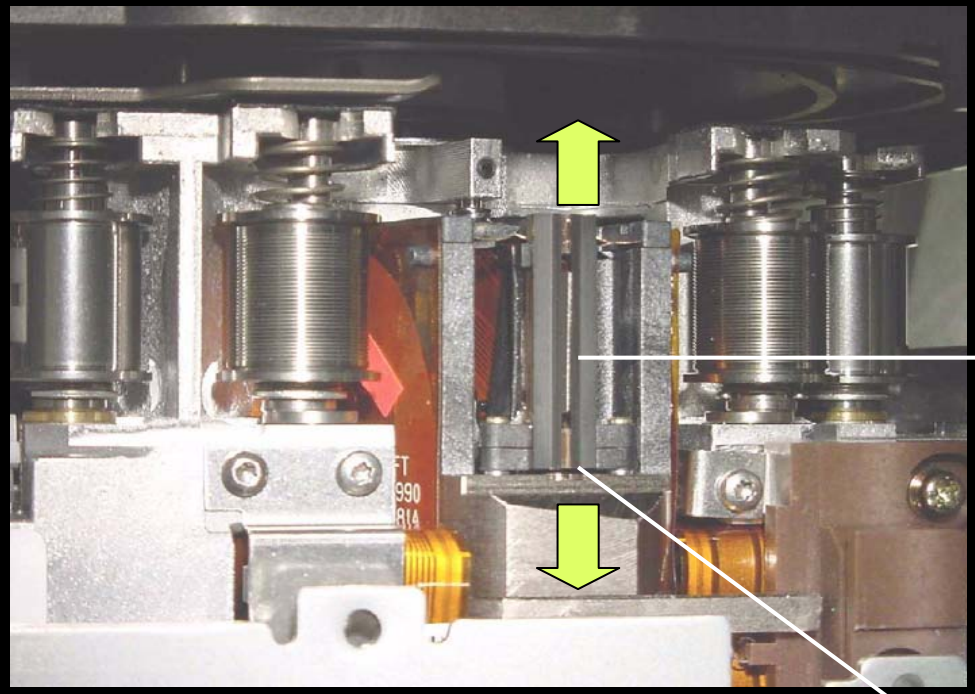
LTO und menschliches Haar



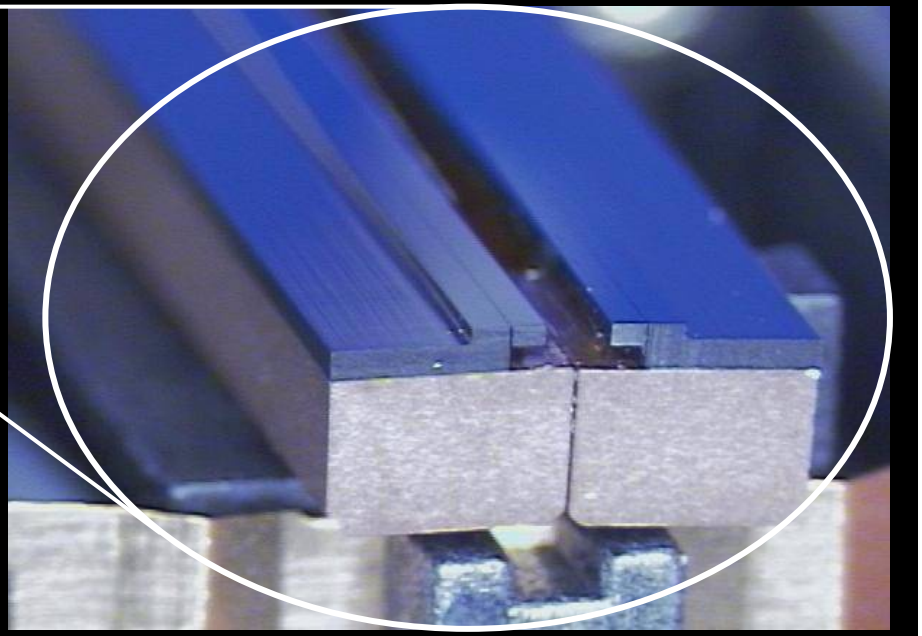
Nutzen von "Write Wide – Read Narrow"



8 TB Herausforderung: Dynamische Spurverfolgung



Feinere Tracks =
schnellere Kopf-Servonachführung
benötigt – bei 4~6 m/sec



Themen

1. Ausblick FC-, SAS-, SATA- Festplatten
2. Flash Technologie
3. Auswege aus dem Festplattendilemma
4. Neue Speichertechnologien

"Festplattendilemma": Viele TB, wenig IOPS

Bessere CACHING Algorithmen

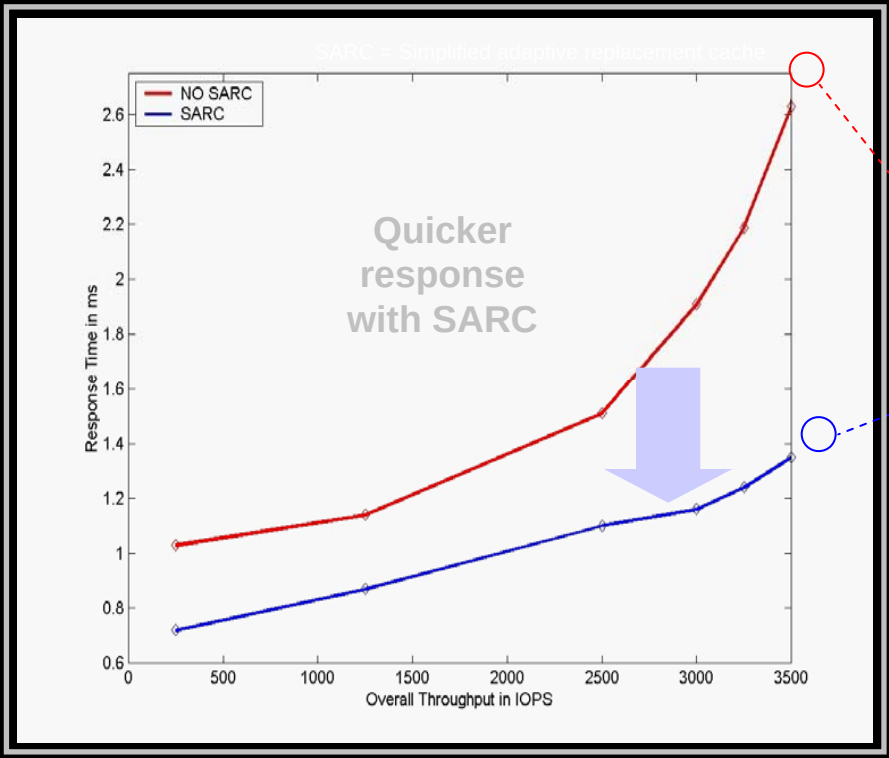
Striping & DECLUSTERED RAID

Bessere Cache Effizienz: SARC (Pat.)

Revolutionäre Strategie für "cache full" Zustand – typisch nach 30~60 min

Industriestandard ist LRU = *purge least recently used*

Neuer Algorithmus SARC = *purge least likely to be re-used*



Effective cache size: +33%

Peak Throughput: +12.5%

Cache miss rate: 11% reduced

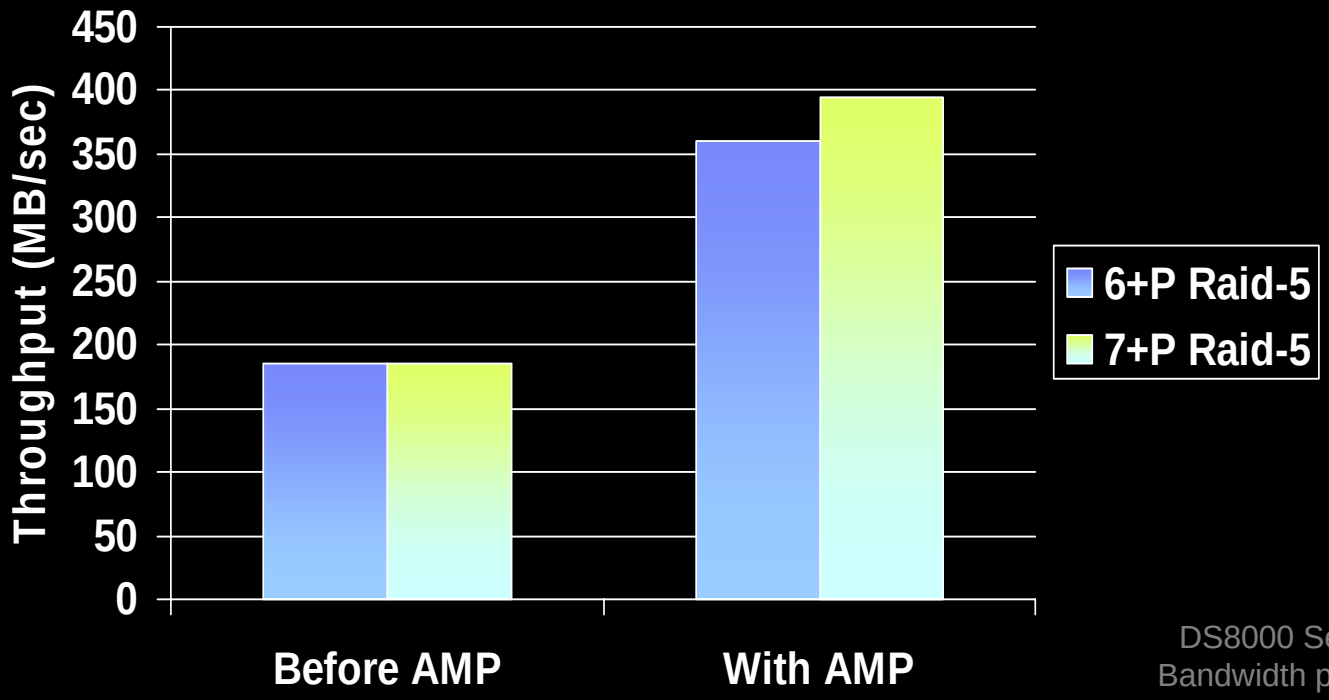
Response times @ 4000 IO/s:
7,60 ms with LRU
1,88 ms with SARC

SARC bereinigt den Cache balanciert nach letztem Zugriff (LRU) und nach Anzahl der Zugriffe (LFU) pro Block

Bessere Cache Effizienz: AMP (Pat.)

Bessere PREFETCH Vorhersage: Adaptive Multi-stream Prefetch AMP

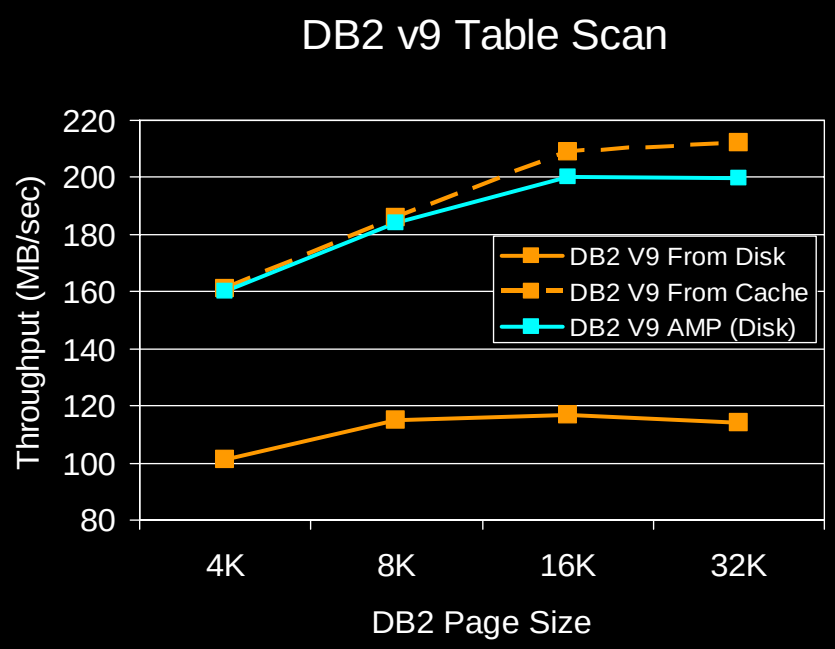
Herausforderung: *n* Clients lesen Daten, mit multiplen VMs pro Client, mit multiplen Anwendungen pro VM… Mustererkennung im Chaos !

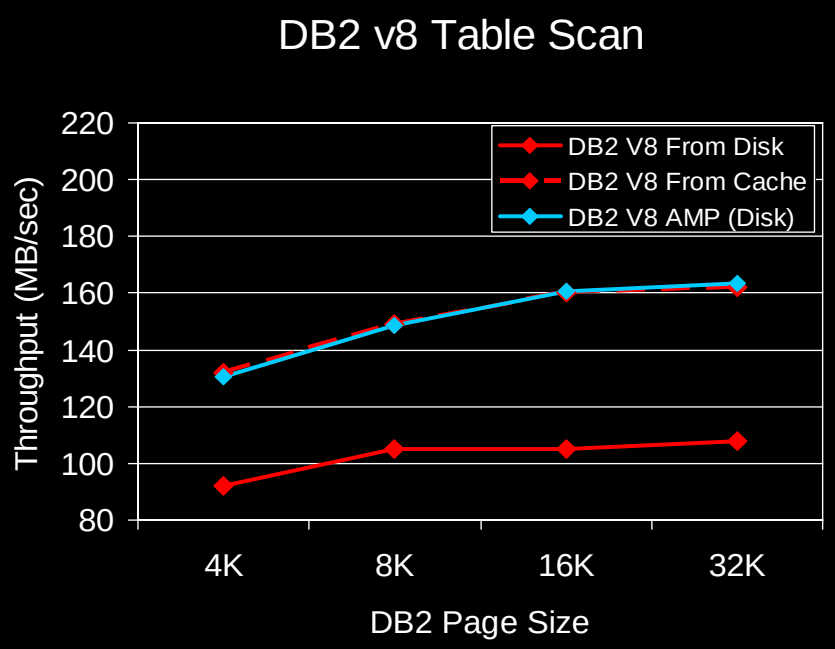


DS8000 Sequential Read Bandwidth per RAID-5 Array

Bessere Cache Effizienz: DB2 Beispiel

DB2 *Table Scans* – mit Cache-Accelerator in SSD-Geschwindigkeit !
(Optimale Prefetch-Vorhersage)





DB2 Table Scan mit AMP
entspricht Table Scan im Cache

Bessere Effizienz = Höhere gefühlte Performance

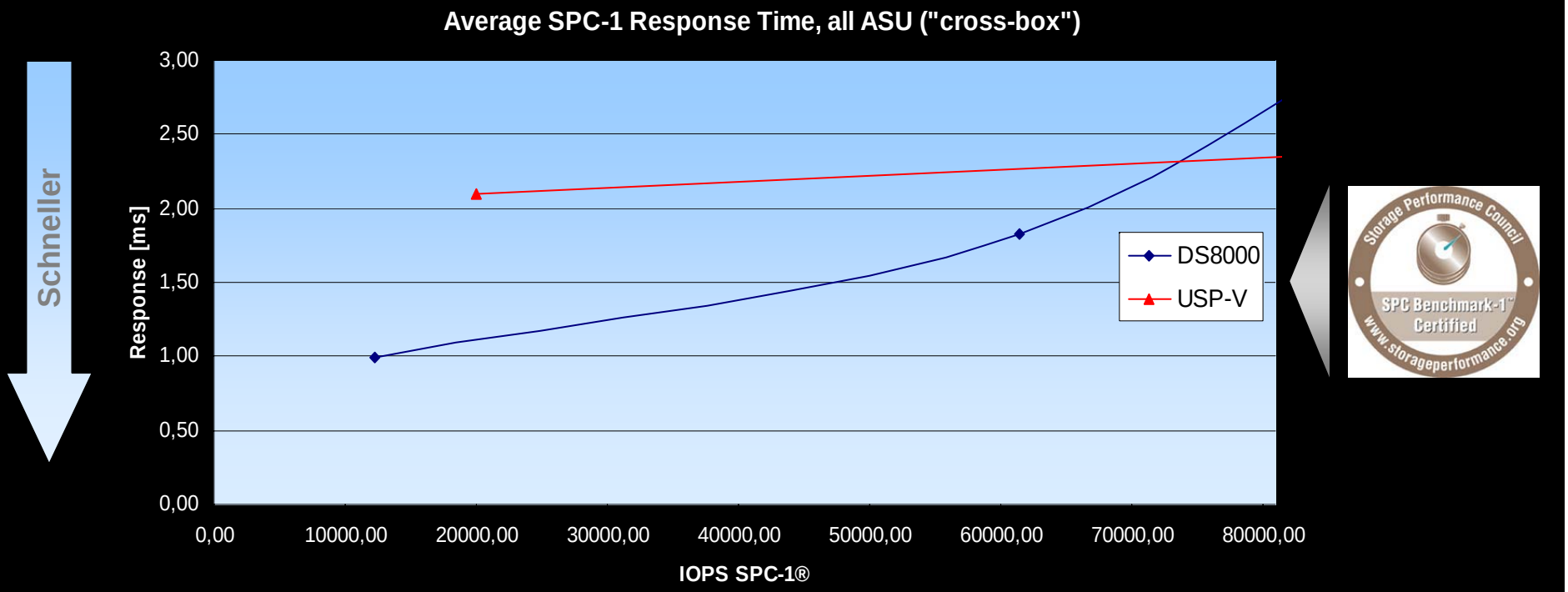
Nur 1 Enterprise Storage Hersteller
liefert eine mittlere Antwortzeit von **1ms**
im SPC-1® Speicherbenchmark… IBM

www.storageperformance.org

Nächster Wettbewerber (Disk-basiert) bei 2 ms

Gefühlte Performance im Normalbetrieb

Benchmark einer zwei Jahre alten DS8000 (R2) mit 512 Platten gegen den derzeit schnellsten Wettbewerber mit 1024 Platten



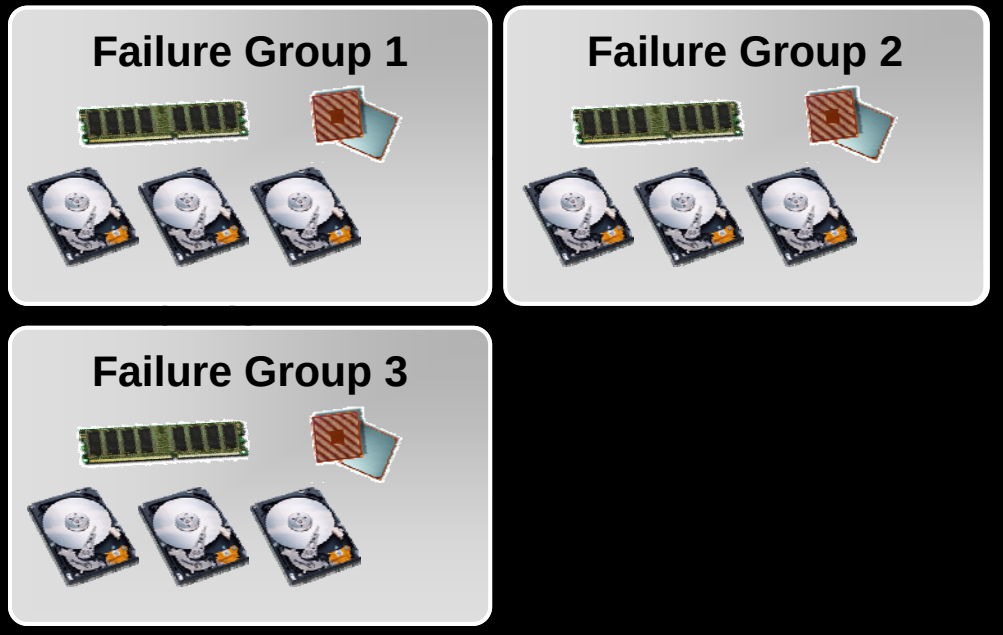
"Festplattendilemma": Viele TB, zu langes *Restore*

Bessere CACHING Algorithmen

DECLUSTERED RAID

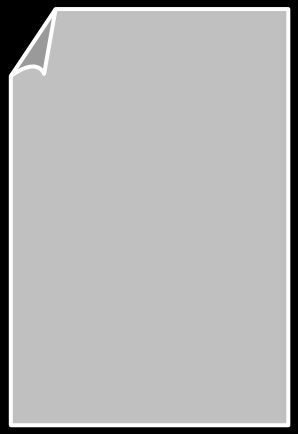
Declustered RAID: Architektur für SATA Speicher

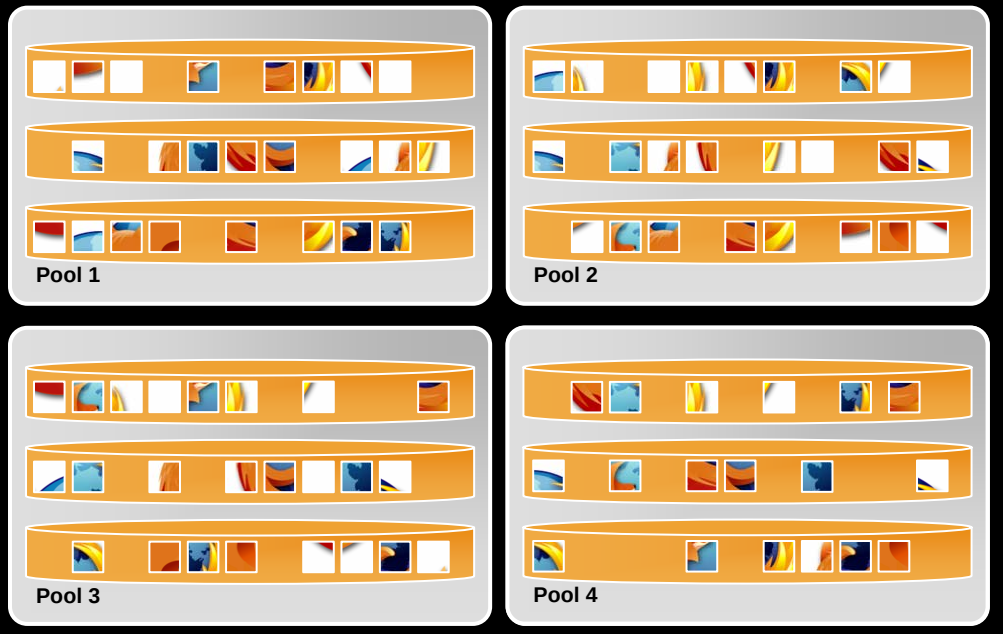
- Unzuverlässige und langsame Platten
- Stochastisches Striping vermeidet IOPS *Hotspots*
- *Failure Groups* enthalten Komponenten, die evtl. zeitgleich ausfallen
- Maßnahme:



Declustered RAID ist transparent erweiterbar

- Ausbau = Mehr Kapazität
- Ausbau = Höhere Performance für alle Daten
- Ausbau = Höhere Verfügbarkeit

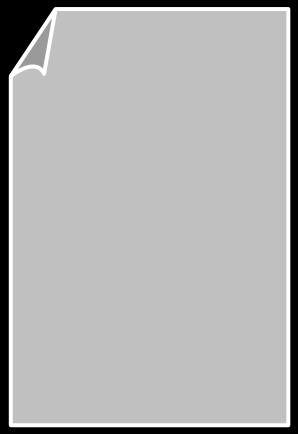


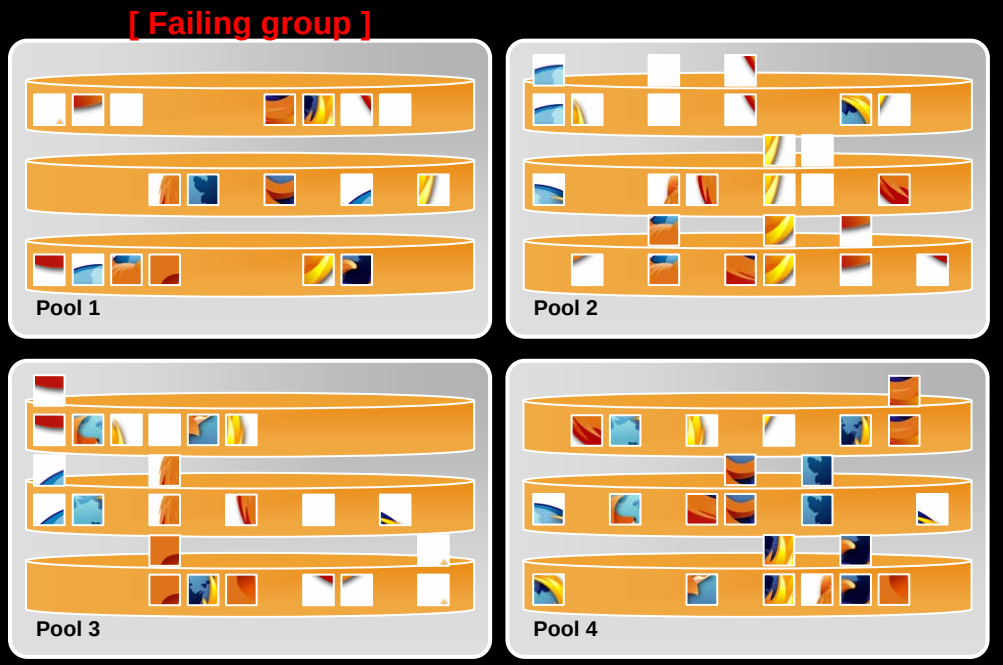


[Adding hardware]

Declustered RAID kommt ohne Rebuild-Prozess aus

- Failure Group Ausfall:
  1. Alle Daten sind noch in anderen Failure Groups vorhanden
  2. Schnelle parallele Wiederherstellung der Duplikate statt *Single Disk Rebuild*
- *Keine Hot Spares*

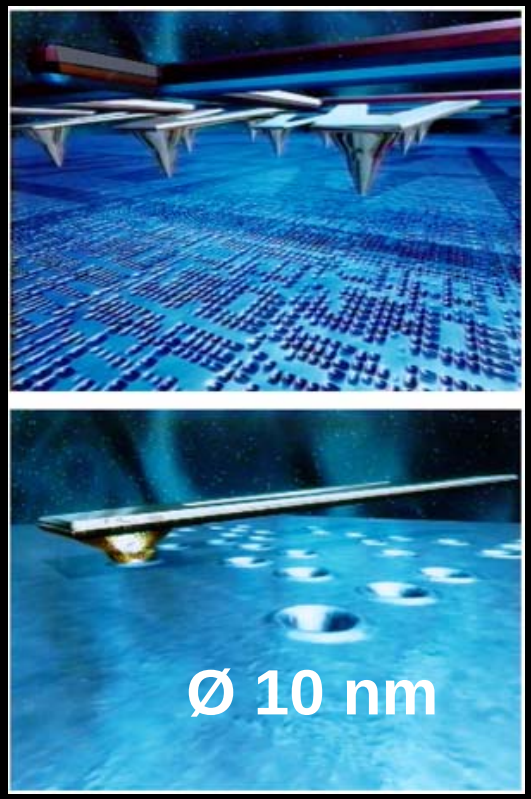




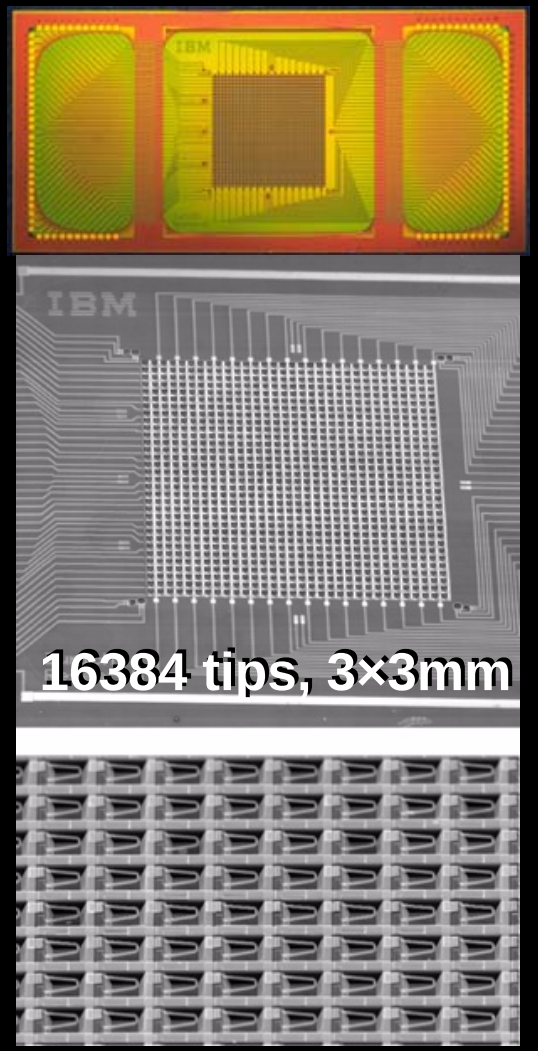
Themen

1. Ausblick FC-, SAS-, SATA- Festplatten
2. Flash Technologie
3. Auswege aus dem Festplattendilemma
4. Neue Speichertechnologien

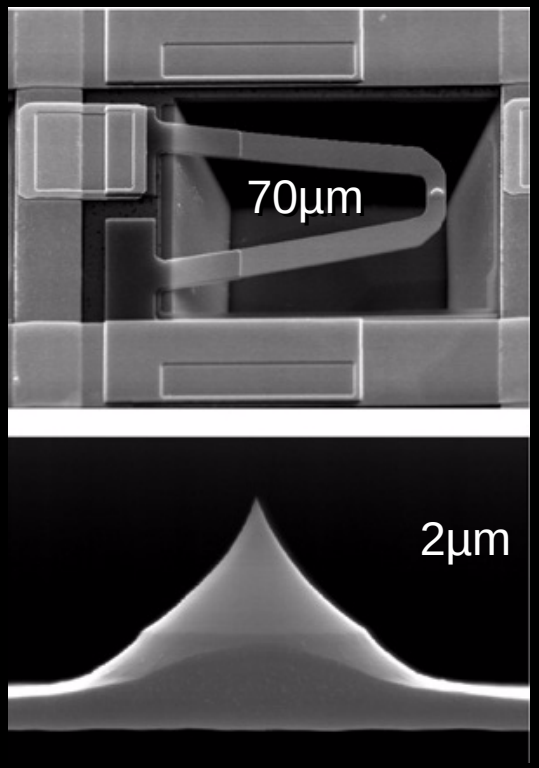
Nanomechanischer Flashspeicher



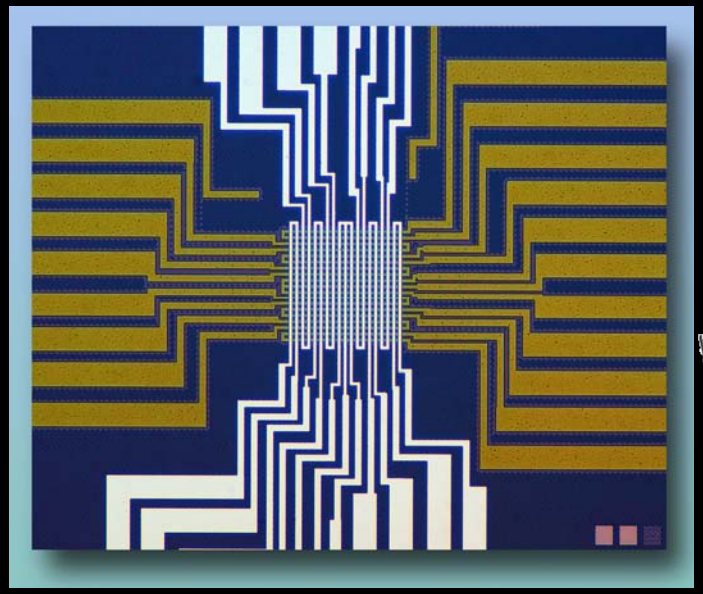
1 Tip liest 1000...2000 Löcher/sec
128x128 Tip Prototyp: 120 Mbit/s
Verbrauch: ~100 milliwatt
Schreibzyklen: >100000
Dichte: 1 Tbit/inch² ++



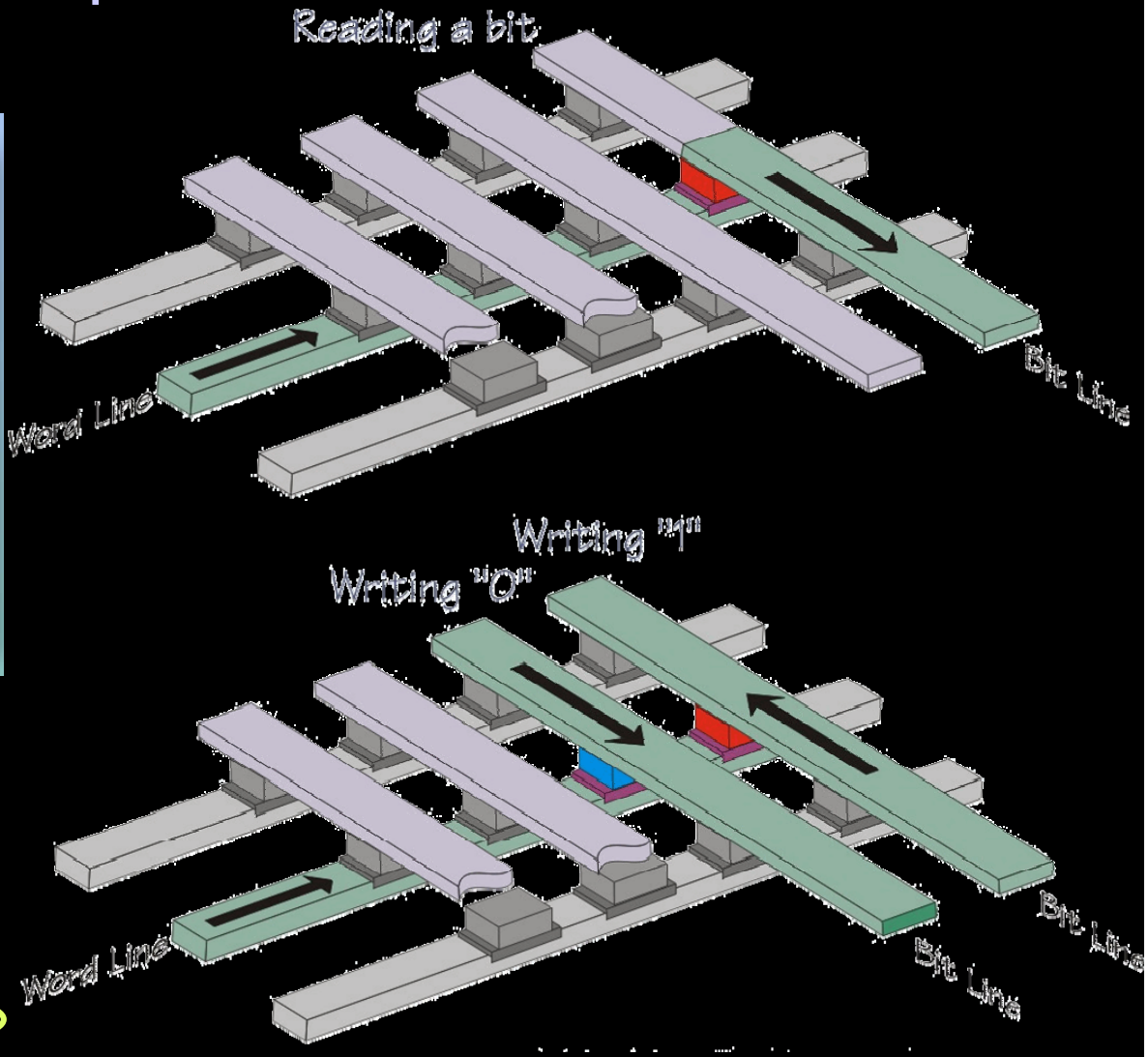
Gerd Binnig von IBM Rüschlikon gewann 1986 den Nobel Preis fürs *scanning tunneling microscope*, die Basis des heutigen "Millipede"



Magnetischer "Flash" Speicher



IBM Prototyp 199..



Produkt von Freescale gefertigt – "MR2A16A" mit 4MBit nonvolatilem Speicher bei 35 nsec Zugriffszeit;

Neuer IBM Demonstrator mit 2ns

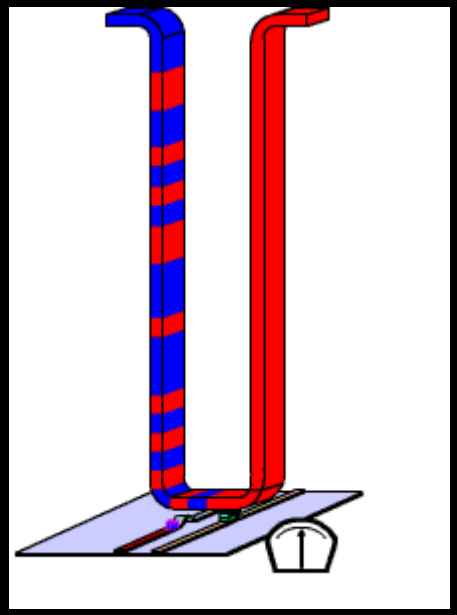
Neueste Entwicklung: "RaceTrack" Speicher

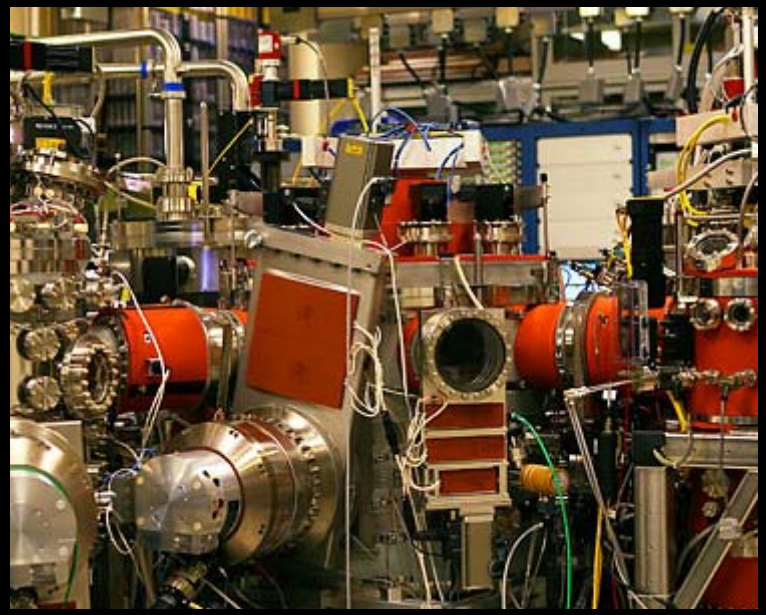
Spintronics

Stuart Parkin, IBM Fellow, Erfinder der GMR Leseköpfe, entwickelt "racetrack memory" auf Spintronics Basis bei IBM Research, Almaden

GMR = Giant MagnetoResistivity

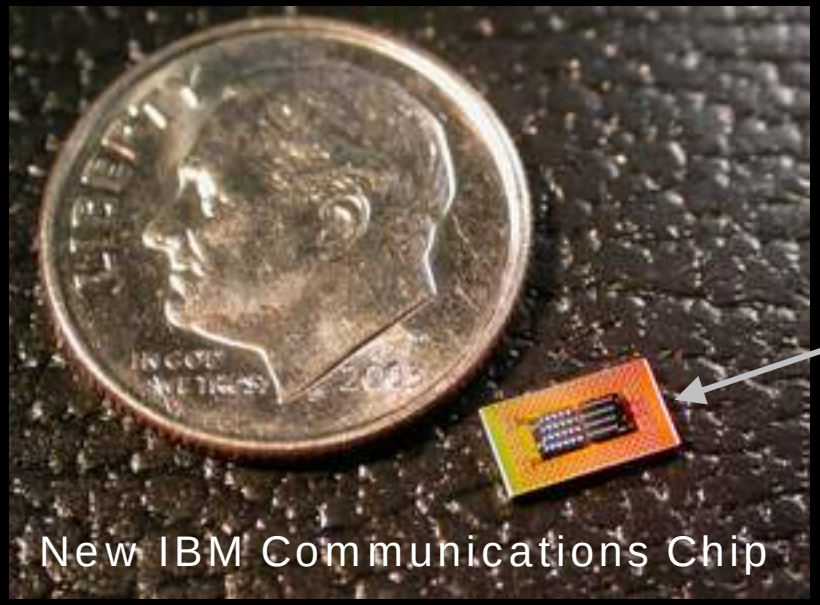
Nobelpreis für Peter Grünberg (D) und Albert Fert (F), 2007





Demonstrator mit 1.5 Gbit/mm²

Die Grenzen von Kommunikationsbandbreite?



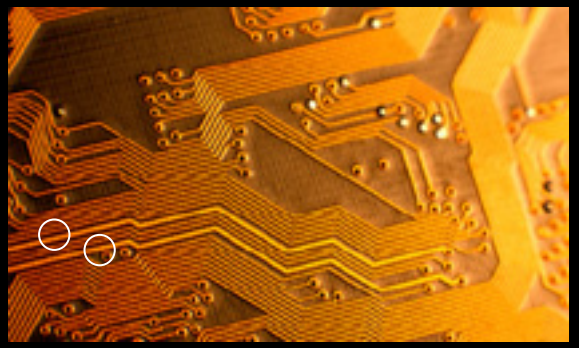
Kombination klassischer CMOS (low power-) Technologie mit Indiumphosphid and Galliumarsenid für optische Teile
Maße: 3.25 × 5.25mm.

160Gb/s: Bandbreite im Überfluss

57,6 Terabyte pro Stunde

HD-DVD Film in 1 sec.

MPEG eines ganzen Lebens (80 Jahre, ½ Petabyte) in 83 h

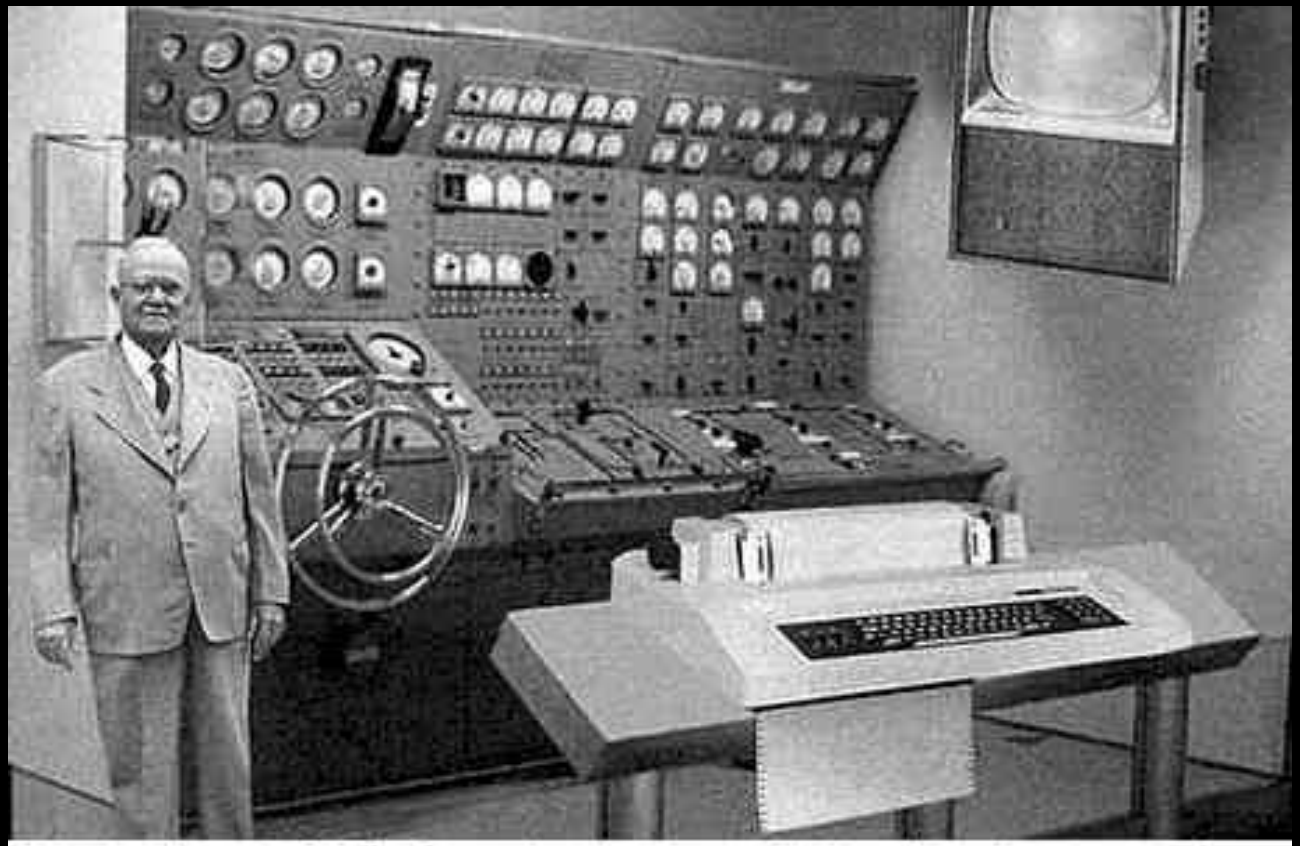


Optische und elektrische Leiterbahnen auf *einem* Träger

März 2007, Optical Fiber Conference, Anaheim

Disclaimer

1954 *Falsche* Vorhersagen: “2004 Home Computer”



Scientists from the RAND Corporation have created this model to demonstrate how a "Home Computer" could look like in the year 2004. However the needed technology will not be economically feasible for the average home. Also the scientists readily admit that the computer will require not yet invented technology to actually work, but 50 years from now scientific progress is expected to solve these problems. With teletype interface and the Fortran language, the computer will be easy to use.

(Anstelle eines Disclaimers: Alle Vorhersagen sind Schätzungen)

axel.koester@de.ibm.com