IBM Systems Storage

Storage Futures
Technology Outlook

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

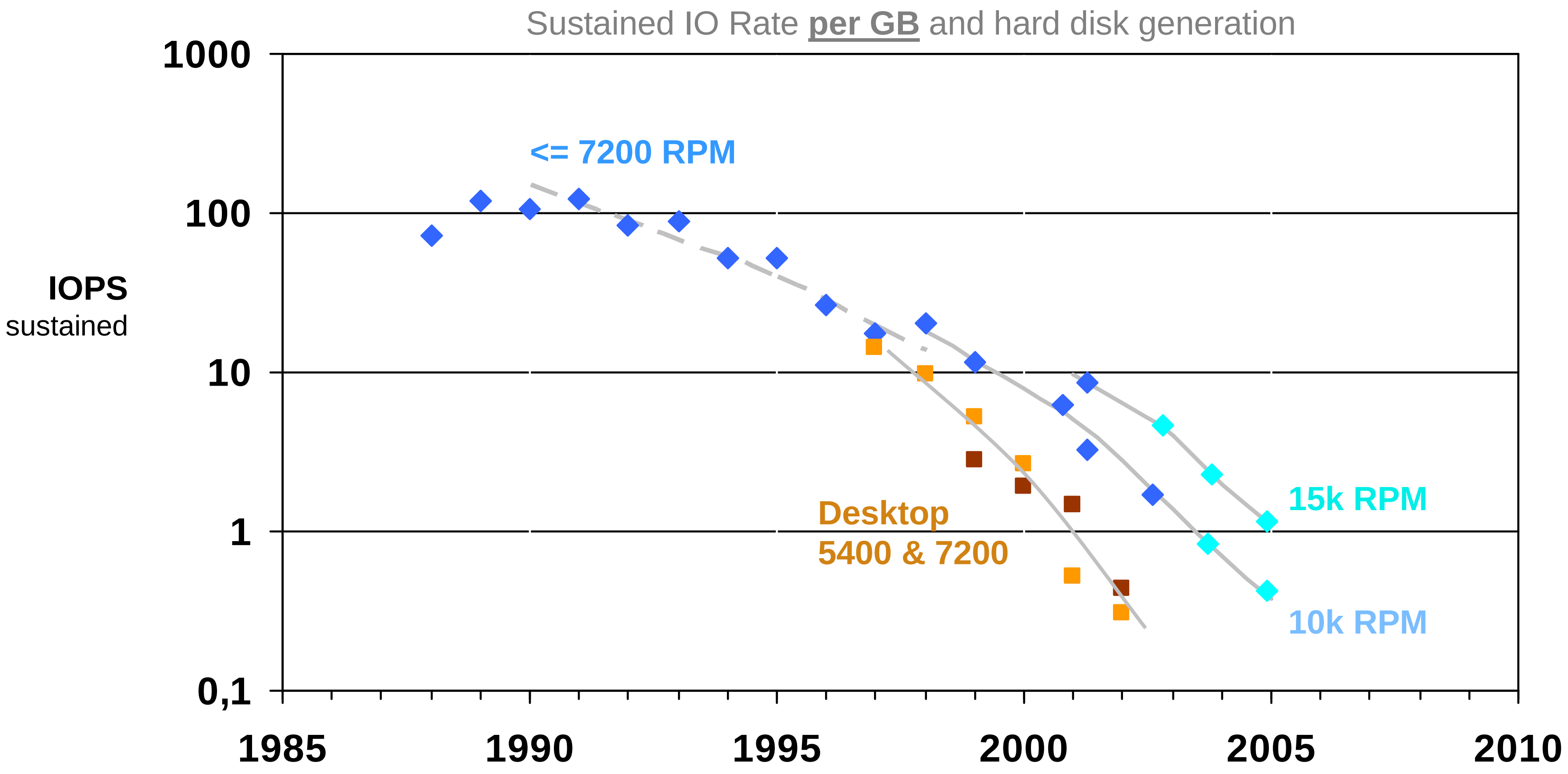
Advances in Caching

Flash Memory

SAN Volume Controller & Flash

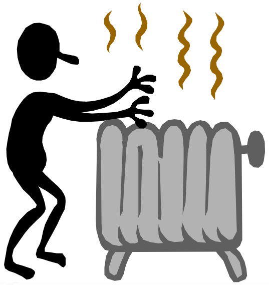
Innovations from the Labs

Hard Disk Performance *per GB* drops alarmingly



Quicker Access? 20k RPM?

- 20.000 RPM disks run **hot**
- **RPM ×2 = Power Consumption up ×8**
- => smaller platters, less capacity

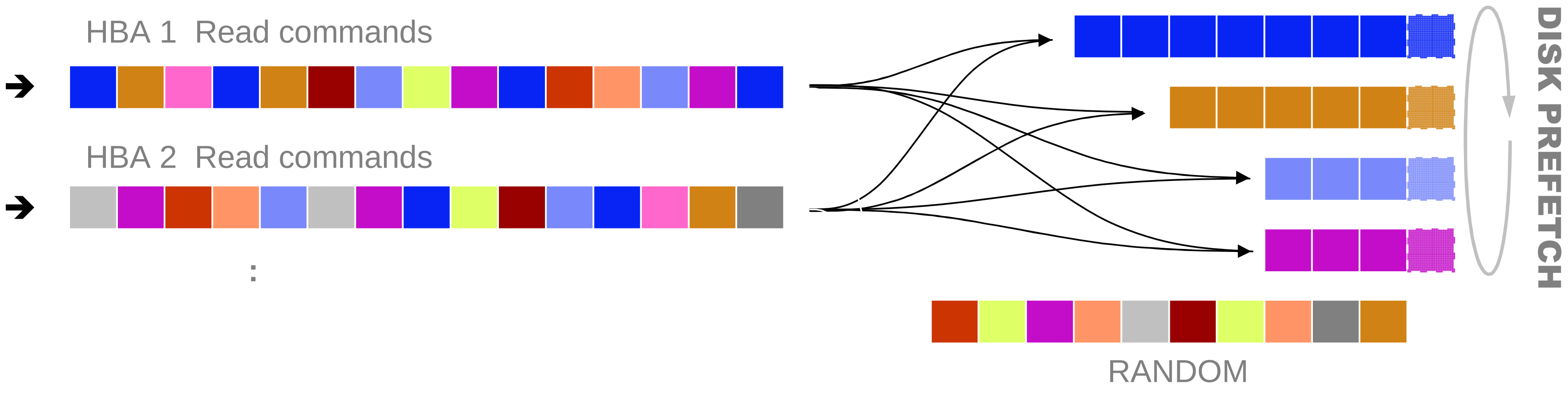


Western Digital ®
VelociRaptor 2.5"
20kRPM Prototype

(*) Air Friction Loss $\sim \{RPM\}^3$

Cache Innovations for larger Disks

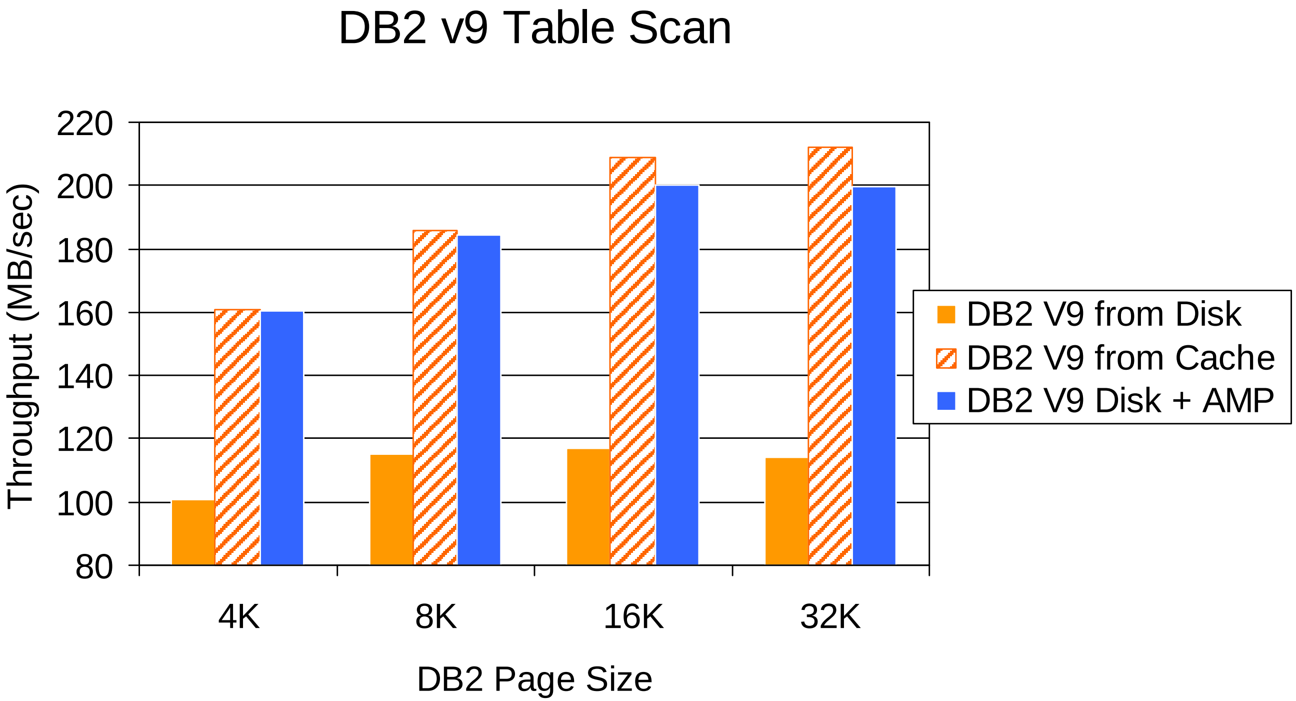
Adaptive Multistream Prefetch Caching



- Finds sequential data patterns in chaotic access streams
- …cross all adapters and IO clients
- …in realtime at > 120.000 IO/s

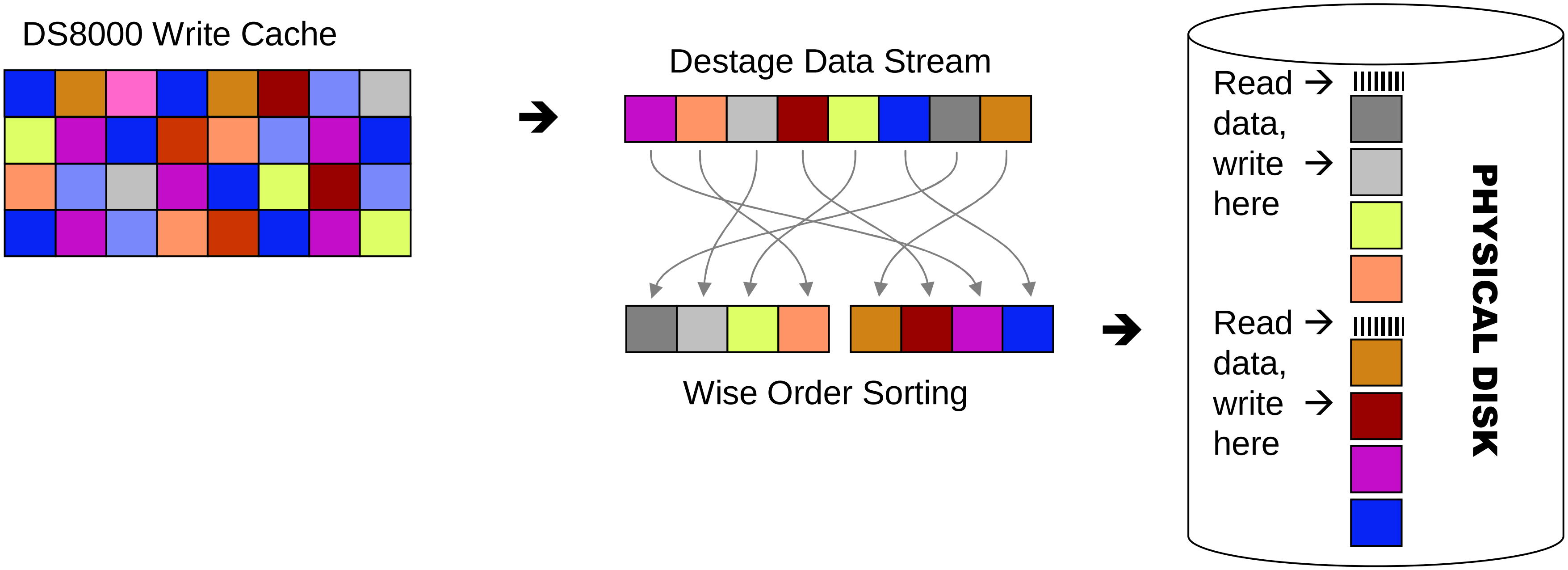
Caching Innovation – DB2 Real World Example

DB2 *Table Scans*, achieving cache speed with SARC & AMP
(= optimal prefetch prediction)



DB2 Table Scan with AMP is equivalent to Table Scan in Cache

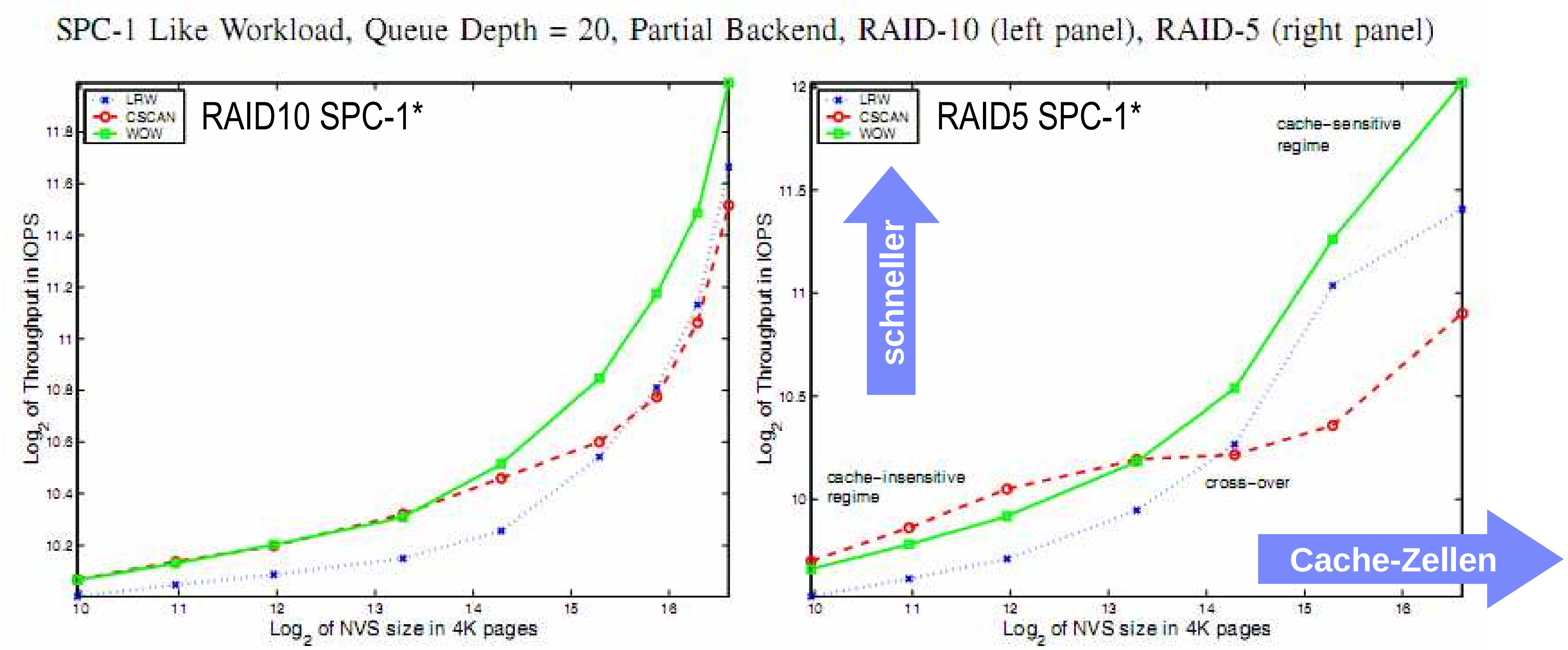
Wise Order Writes (upcoming in DS8000)



- Optimize head movements for low-latency read access
- Delay writes in cache until head is in proximity

➔ much better mixed workload behavior (close to 100% read)

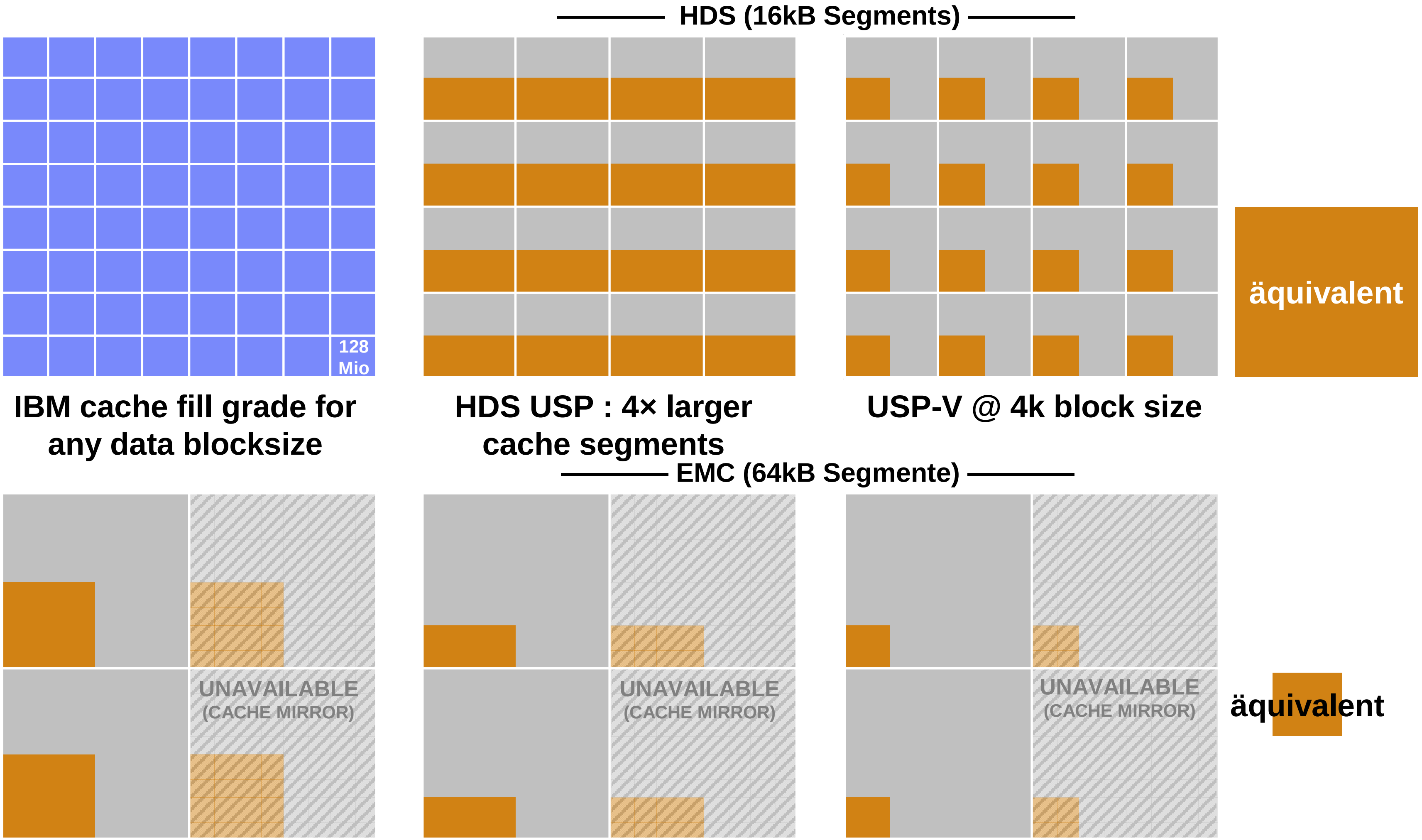
Wise Order Writes compared with "classics"



- **WOW** *(wise order writes)* versus "second best" **CSCAN** *(cyclical scan)* and classic **LRW** *(least recently written)* under SPC-1-like workload

(*) **simulated** SPC-1® OLTP Workload; R/W = 40/60, each 60% random

Cache Segmentation : DS8000 / SVC and competitors



HDS (16kB Segments)

128 Mio

IBM cache fill grade for any data blocksize

HDS USP : 4× larger cache segments

USP-V @ 4k block size

äquivalent

EMC (64kB Segmente)

UNAVAILABLE (CACHE MIRROR)

UNAVAILABLE (CACHE MIRROR)

UNAVAILABLE (CACHE MIRROR)

äquivalent

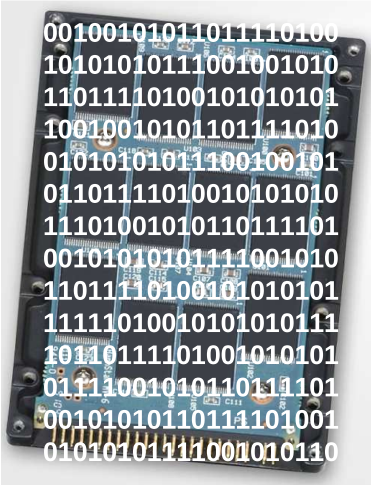
Flash Memory versus (?) Disks

Internal Flash Performance is non-trivial

Read

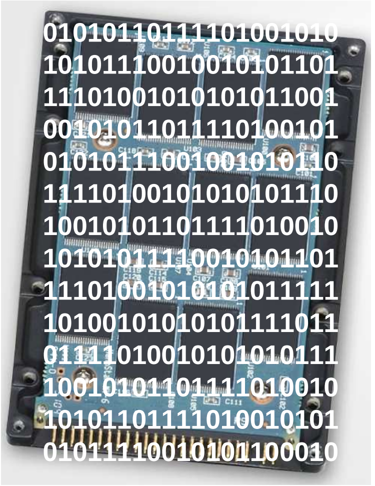
Very fast

Sequential Write



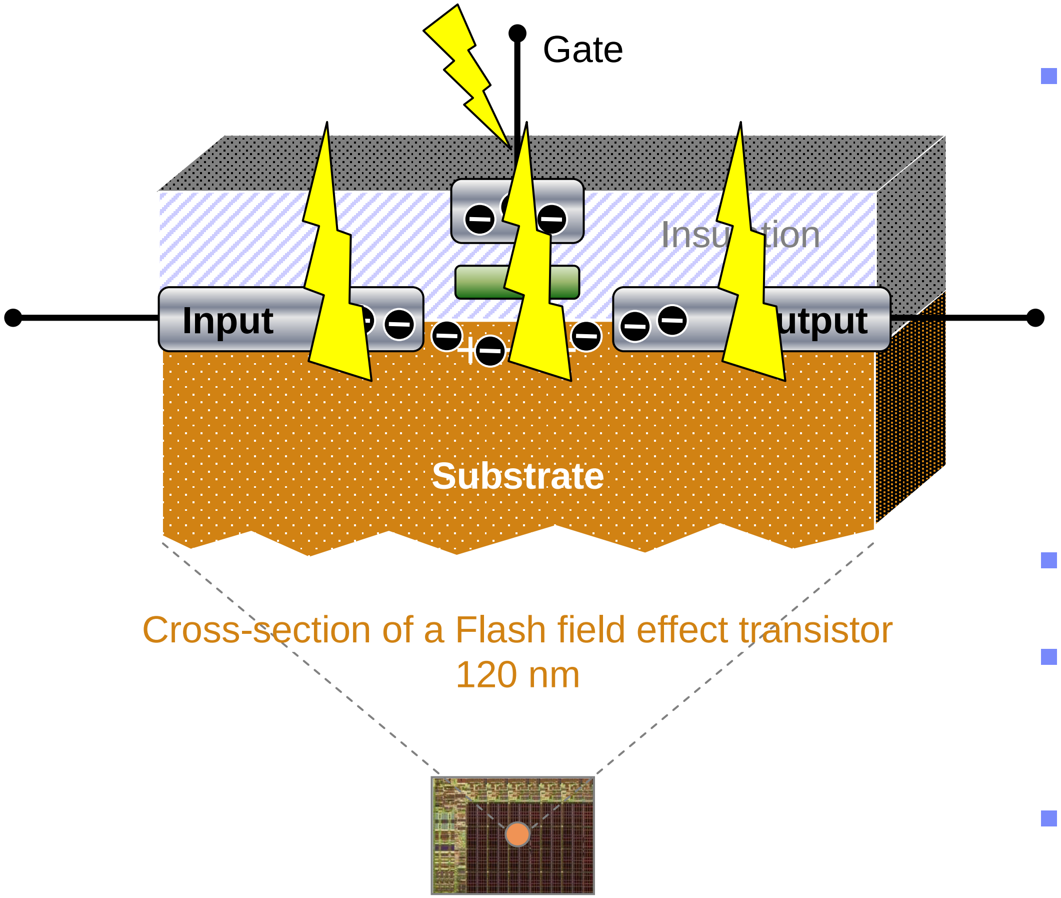
Fast

Random Write



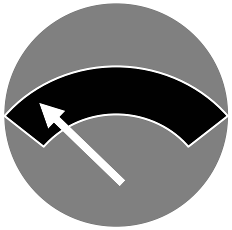
Slow

How Flash Storage Cells work



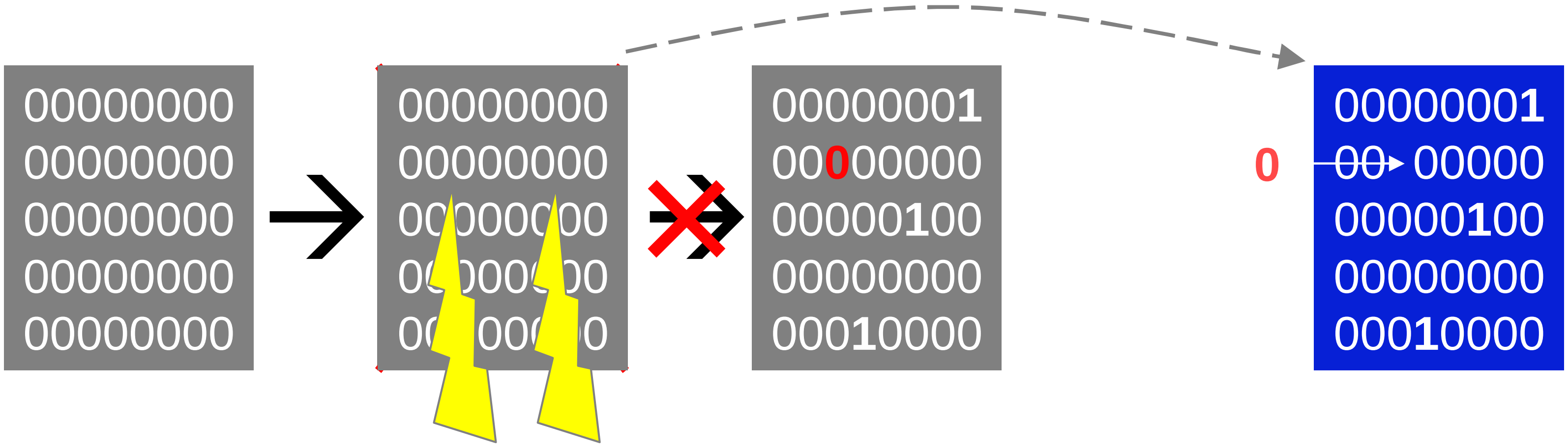
Cross-section of a Flash field effect transistor
120 nm

- **Floating** (insulated) **Gate**



- Only able to write ONEs
- Cannot delete ONEs individually, only block-wise
- Deletion wear after 100.000×

Overwriting & Deleting Flash Data



00000000
00000000
00000000
00000000
00000000

00000000
00000000
00000000
00000000
00000000

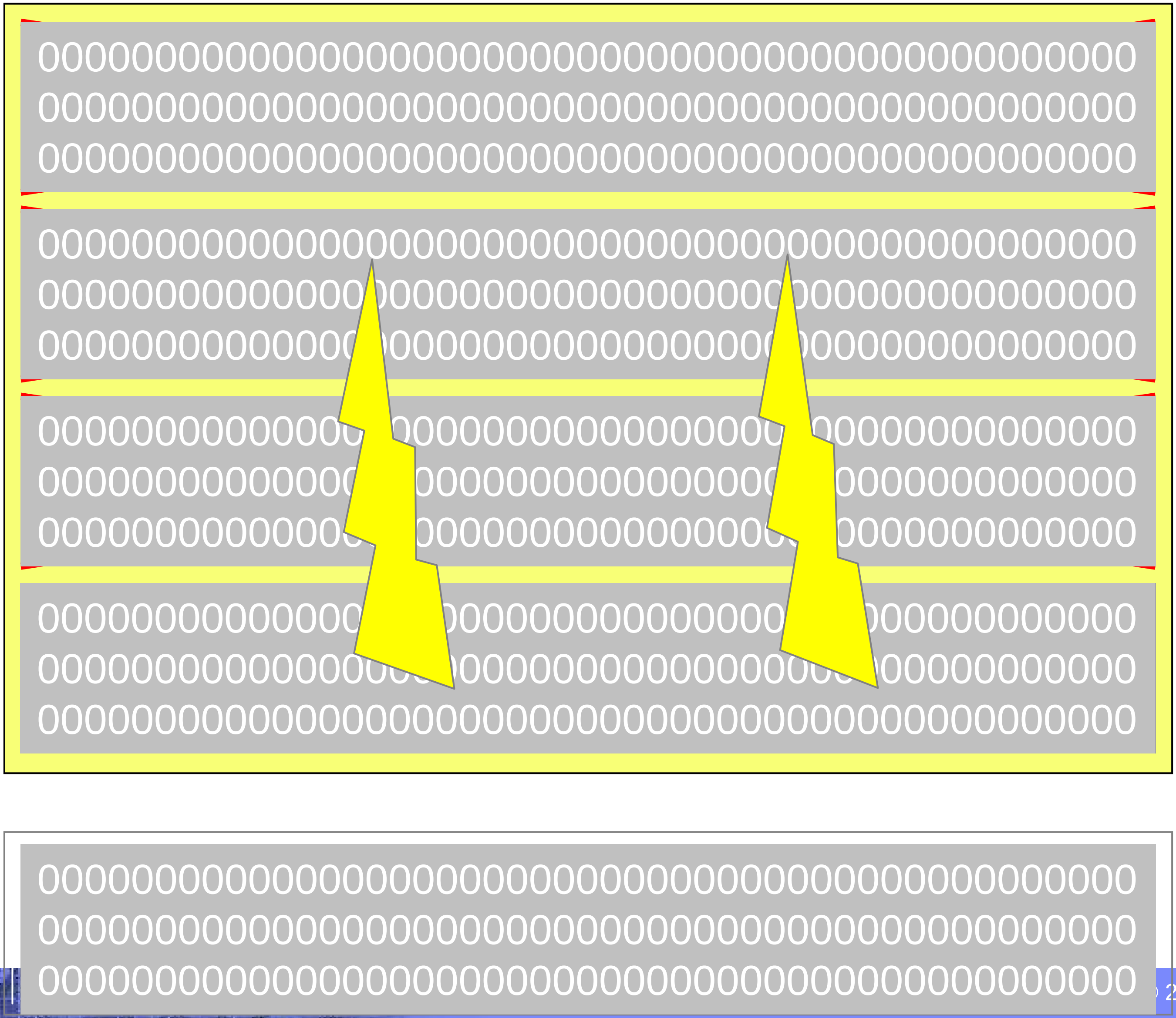
00000001
00000000
00000100
00000000
00010000

0

00000001
~~00~~ 00000
00000100
00000000
00010000

- Random Write is *not* the optimal workload for Flash
- **Delete block** sizes are much larger than typical IO size
- Constant relocation of often-used blocks helps leveling wear

How Solid State Drives (SSDs) handle block deletion



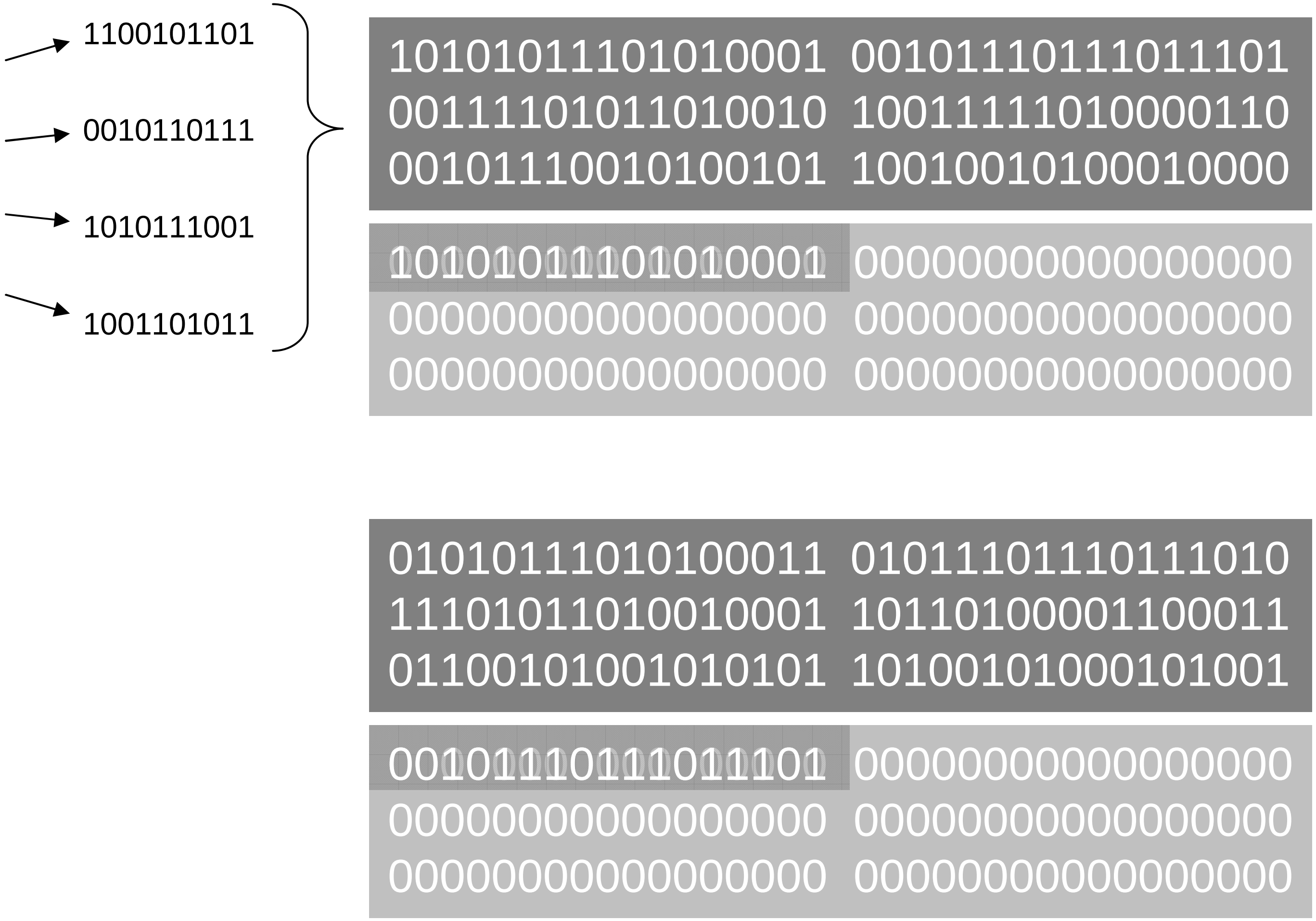
Cost-effective Approach to Random IO Performance

1. Use **Flash PCI-Memory** instead of **SSDs**
2. **Serialize** all random IOs (= less deletions)
3. Don't overwrite in-place
4. Optimize IO patterns at the system level, upfront of the drives

ADTRON® SSD 2,5" 160GB
$80-$115 *per gigabyte*

PCI-based Flash Memory
$30 *per gigabyte*

Random IO Serialization



1100101101

0010110111

1010111001

1001101011

10101011101010001 00101110111011101
00111101011010010 10011111010000110
00101110010100101 10010010100010000

10101011101010001 00000000000000000
00000000000000000 00000000000000000
00000000000000000 00000000000000000

01010111010100011 01011101110111010
11101011010010001 10111010001100011
01100101001010101 10100101000101001

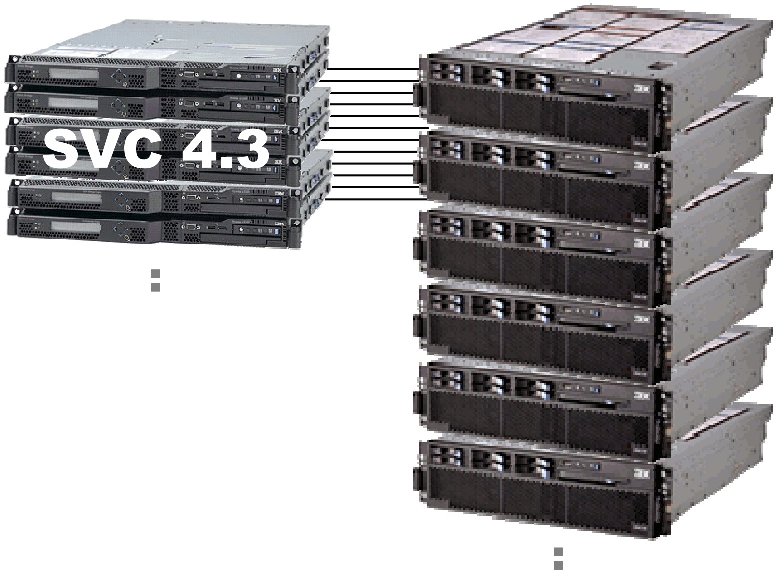
00101110111011101 00000000000000000
00000000000000000 00000000000000000
00000000000000000 00000000000000000

IBM Quicksilver : 1 Mio IOPS World Record

- Technology Demonstrator: IBM SAN Volume Controller + Flash

- Database workload 70/30, 0% cache, running for 2 hours, delivered **1 Mio IOPS** at **700µs response time** (peak 1.1M)

- Flash-optimizing SVC controller →

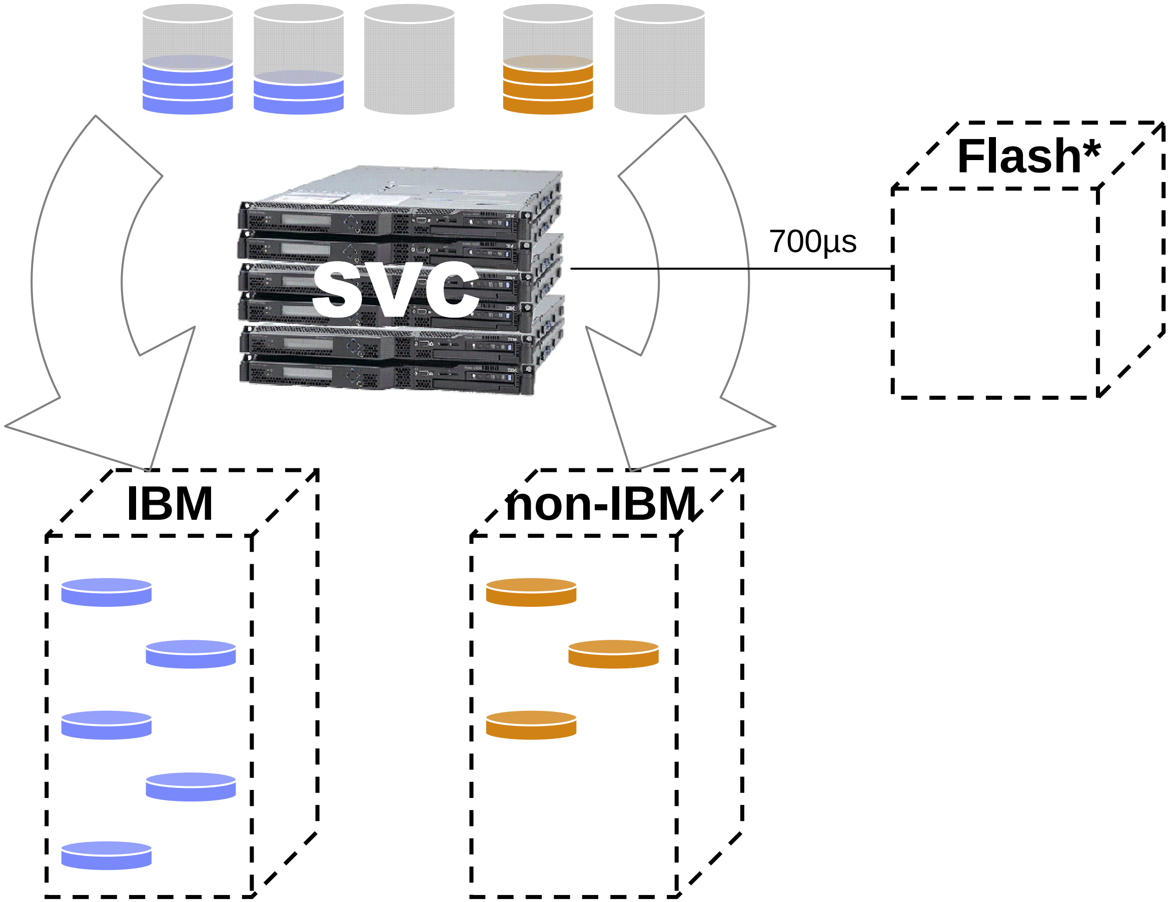


PCI Flash memory in modified SVC cages

August Press Release

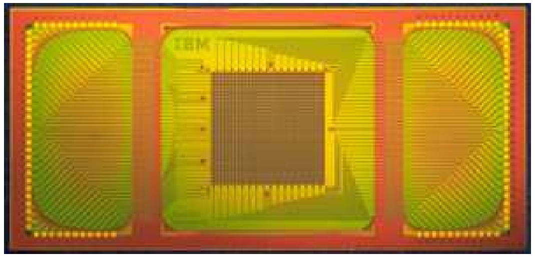
IBM Quicksilver Technology

- SVC 4.3 adds fine-grained "thin provisioning" to *any* storage

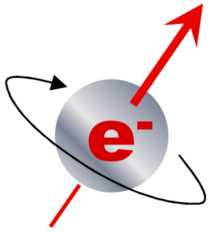


- The smallest SVC grain size is 32kB
- Maximizes the use of your storage

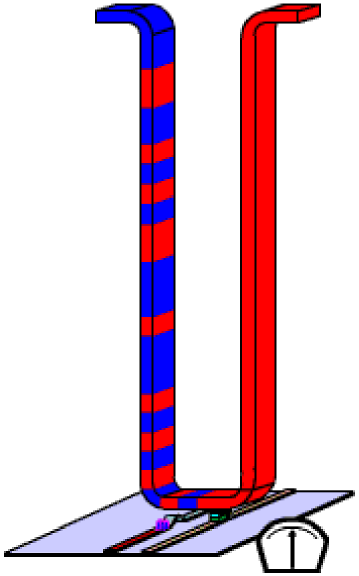
"Solid State" Alternatives

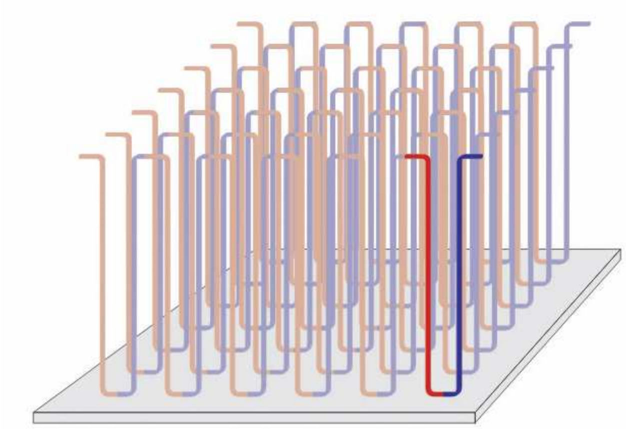


Spintronics & "RaceTrack" Memory



Spintronics

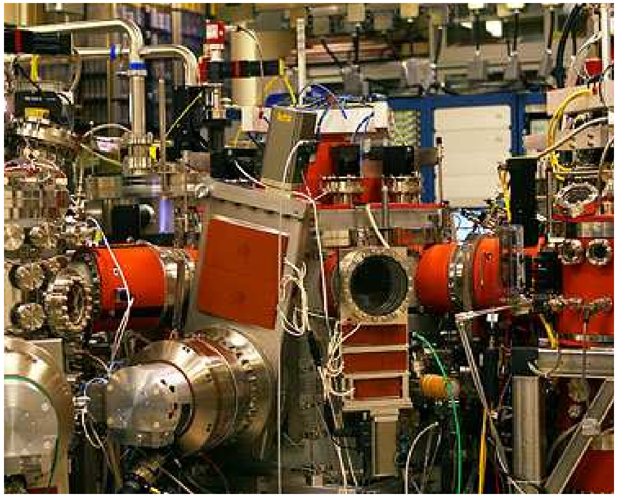




RaceTrack Storage Array

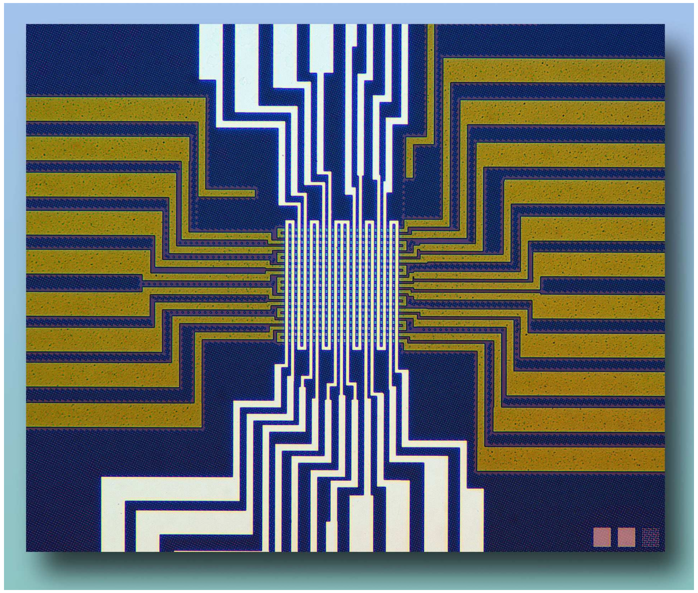
Storage in 3rd Dimension

"Large" read/write head, "small" bits on ferroelectric nano wire



IBM Fellow Stuart Parkin,
Inventor of GMR read heads,
investigates "Racetrack Memory"

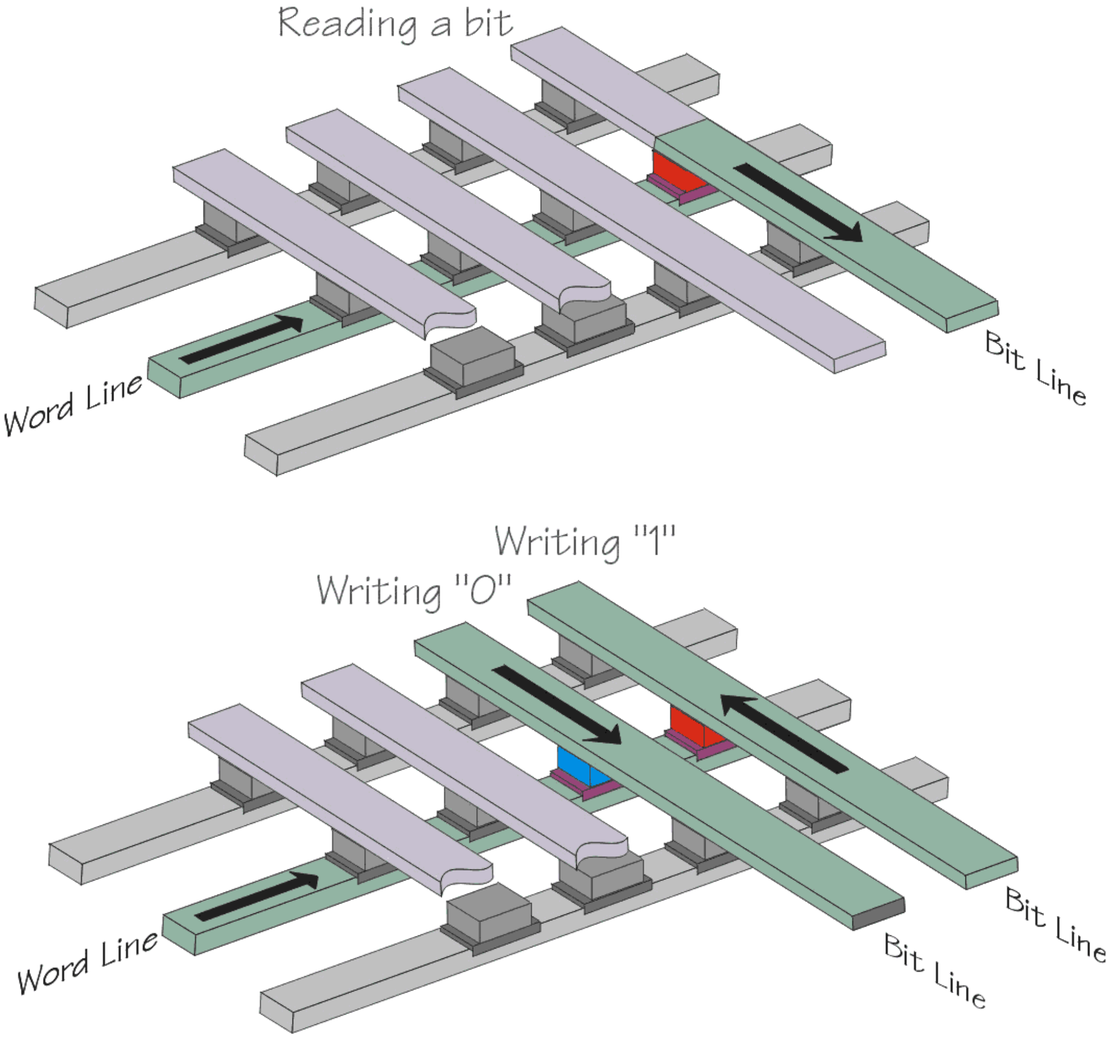
Magnetic Random Access Memory



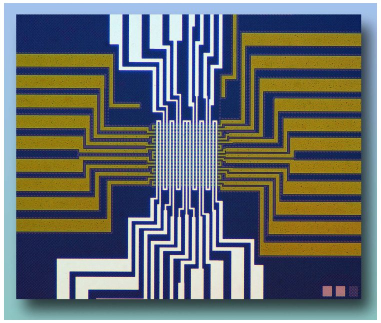
IBM Prototyp 199..

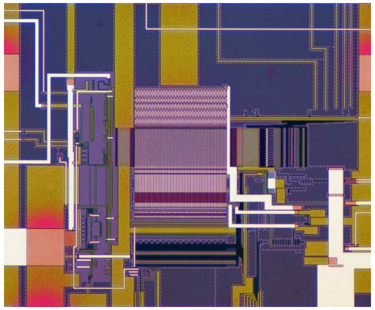
Manufactured @Freescale Inc.
"MR2A16A" 4Mb non-volatile
35 nsec Access Time

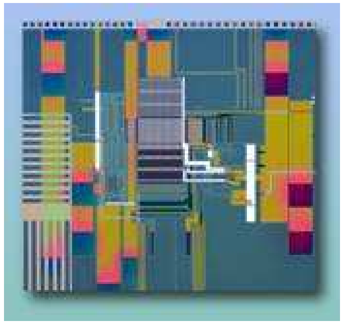
New IBM Demonstrator:
2 nsec Access Time



Non-volatile RAM = IT Revolution !

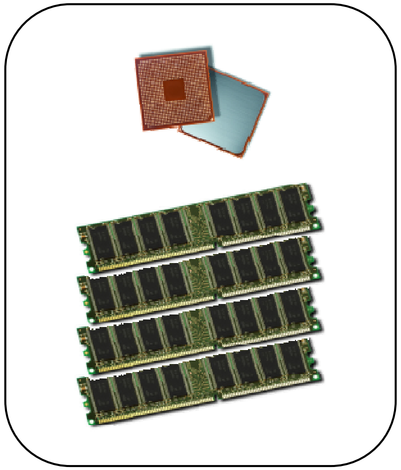




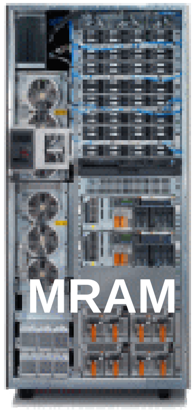


- Magnetic RAM
- Phase-Change RAM
- Ferroelectric RAM
- …

How to use non-volatile RAM



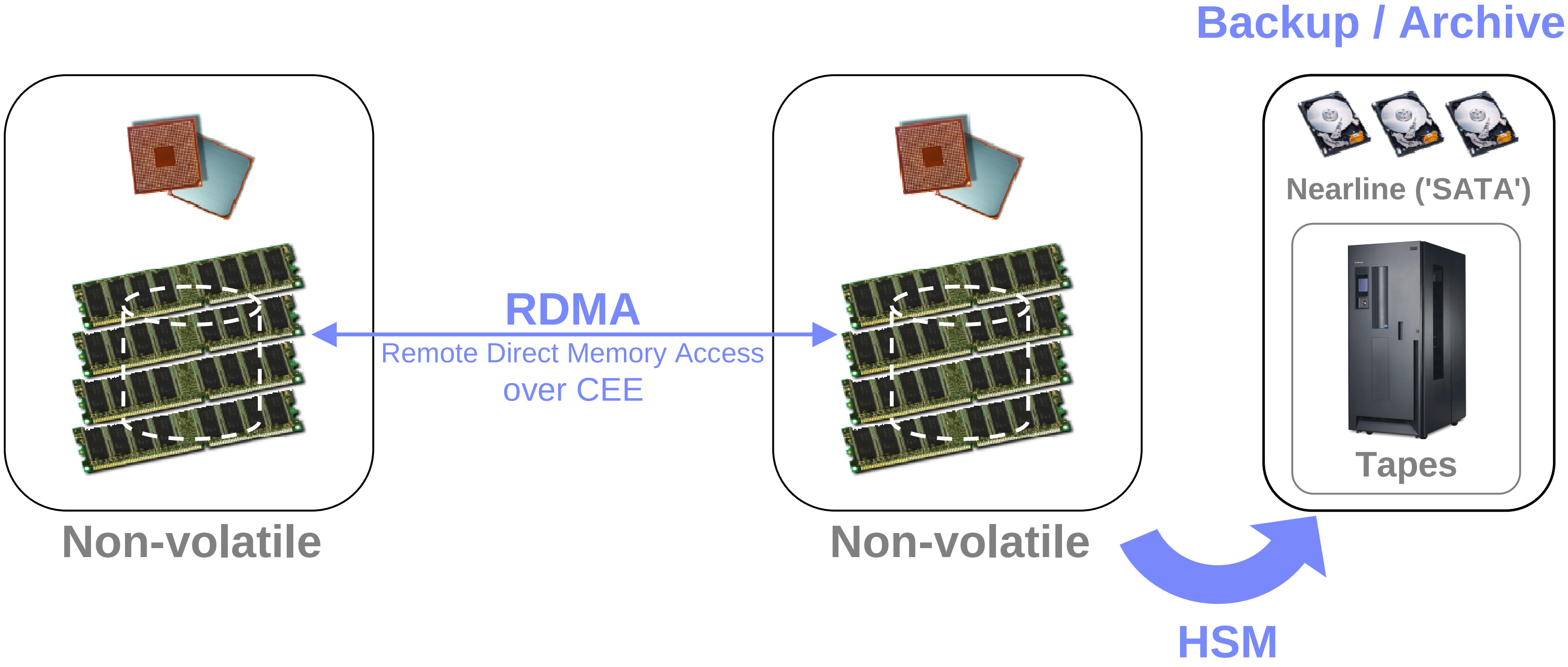
20 nsec



2 nsec

- Online storage moves nearby the processor
- Fast (Fibrechannel-) disks disappear
- Monolithic design = higher reliability

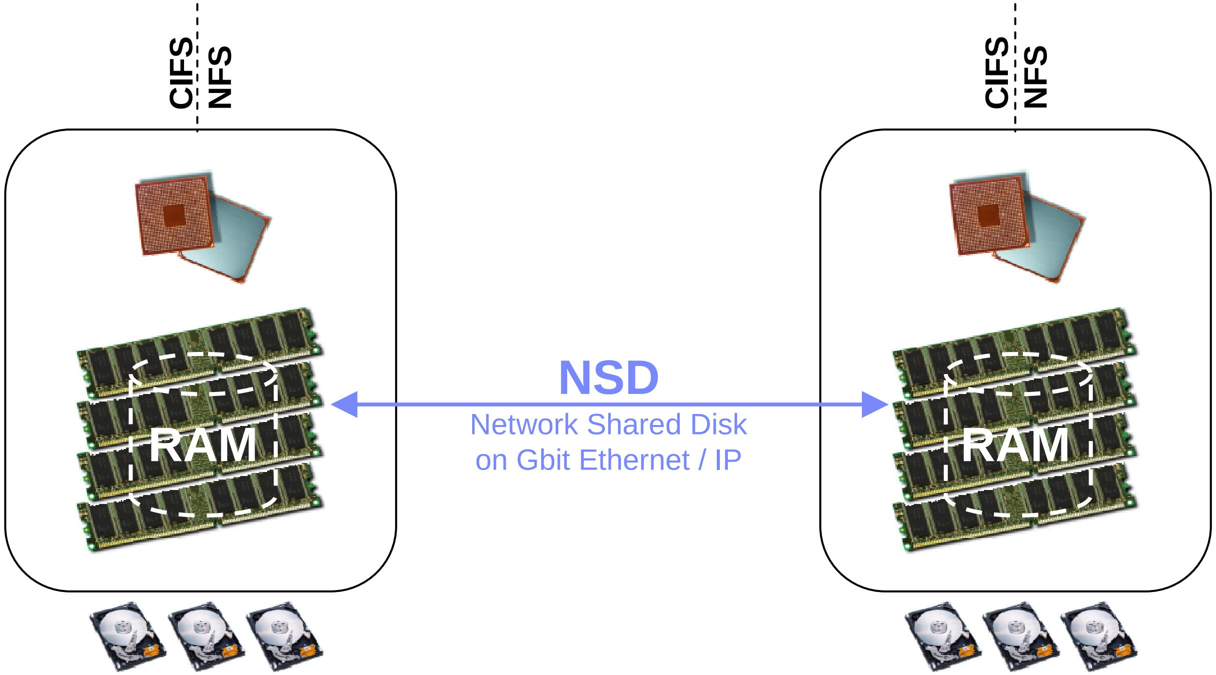
The SAN of the Future



- Memory-to-memory SAN *(RDMA over Converged Enhanced Ethernet)*
- Disks → RAM
- Tapes → Disks
- Paging → HSM

Today: Consistent Caching without non-volatile RAM

Cache as a Hard Disk Memory Extension in a NAS Grid

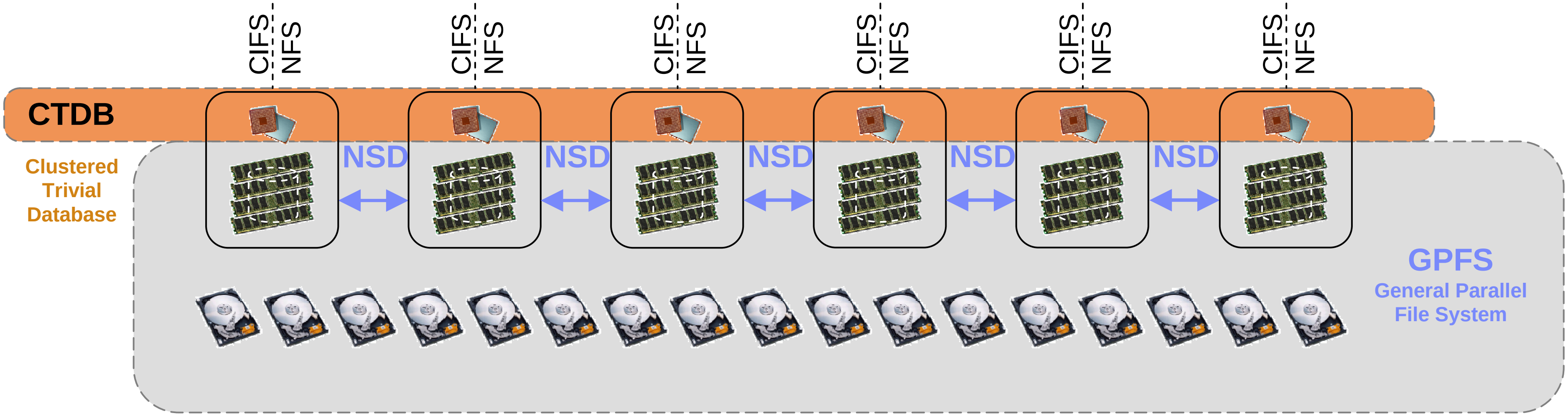


SKIP

Scale-out File Services

As close as possible to "full parallel" in RAM

- Native Samba on a Cluster File System = NO NO !
- For **parallel NAS for Windows**, use CTDB !



- Ultra-fast System Backup: TSM Scan **@1 Mio Files / sec**
- Ultra-fast "virtual" Full Restore

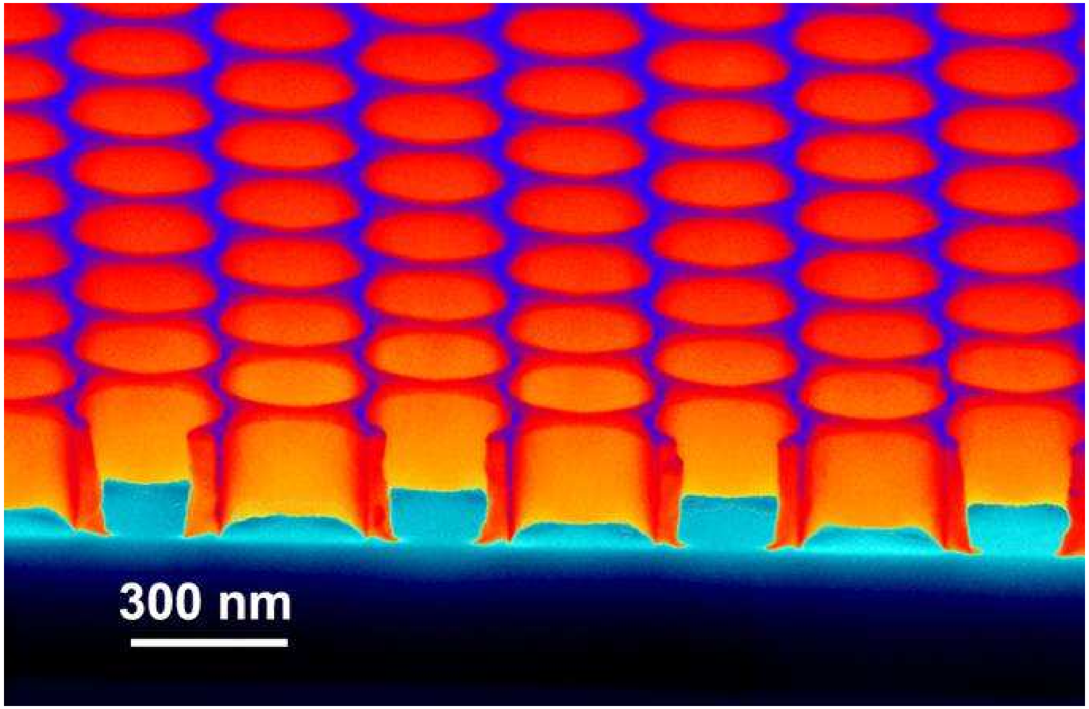
Most Recent Labs News

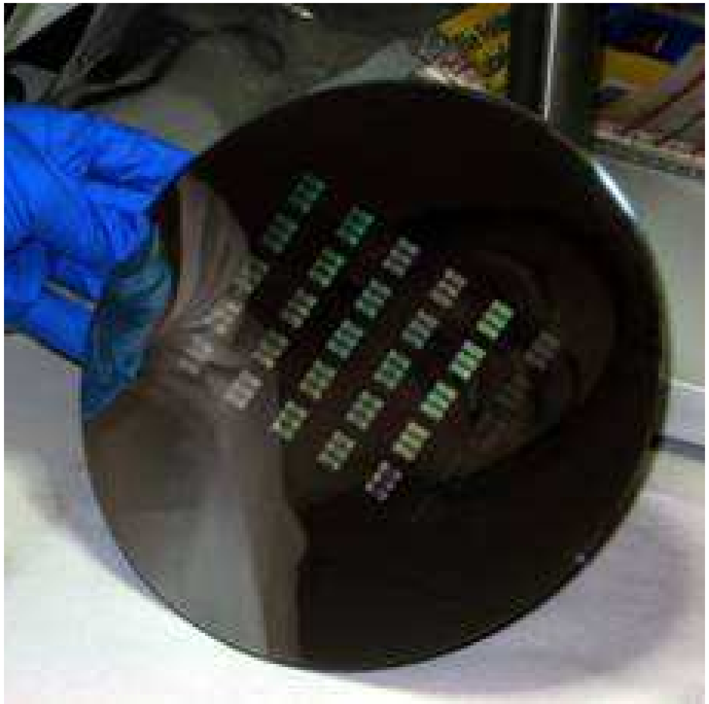
Image © DOCB

Slow Light

nature
SILICON ENHANCED
A chip to control the speed of light
THE FIRST STARS
Spectres in the cosmic infrared background
HURRICANE SEASON 2005
Reporting from the eye of the storm
THE MALARIA PARASITE
Protein networks reveal drug targets
NATUREJOBS
The Italian jobs

- Nano-structured silicon with **refraction index 300**
- Light is 300 times slower (can be influenced)



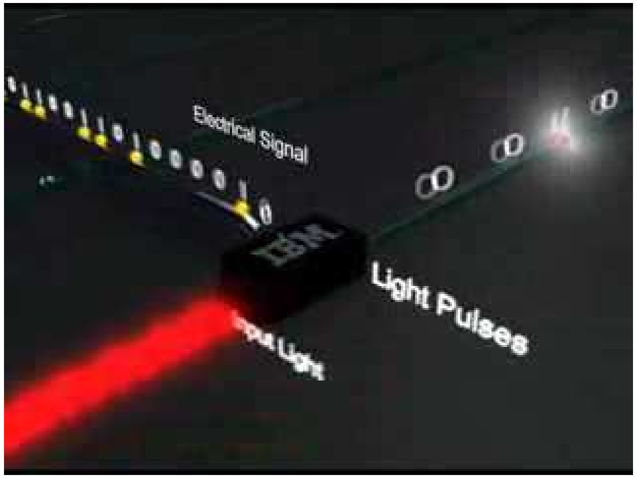


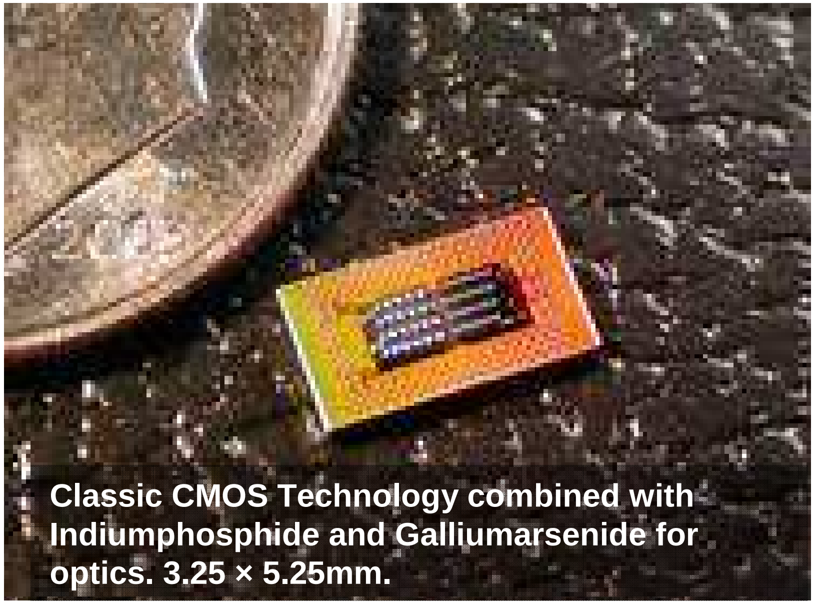
Predictive Summary

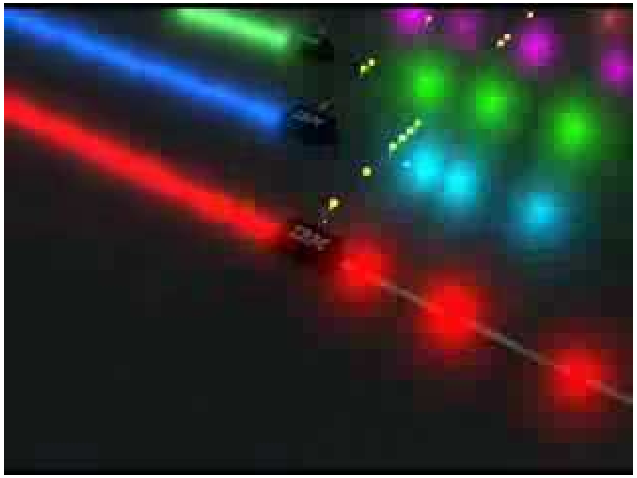
Σ

Future Computers will contain…

- "Slow light" storage *(maybe)*
- 3D nano structures
- Spintronics *(for sure)*




Classic CMOS Technology combined with Indiumphosphide and Galliumarsenide for optics. 3.25 × 5.25mm.



Slow Light : Beware of the Consequences…

"Slow Glas"

IBM

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.