IBM Systems Storage

Storage Innovation

New Technology for Smarter Datacenters

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

27.10.2009

Will Storage continually get cheaper?

Will Flash Memory replace Disks?

Innovations for DS8000 R5

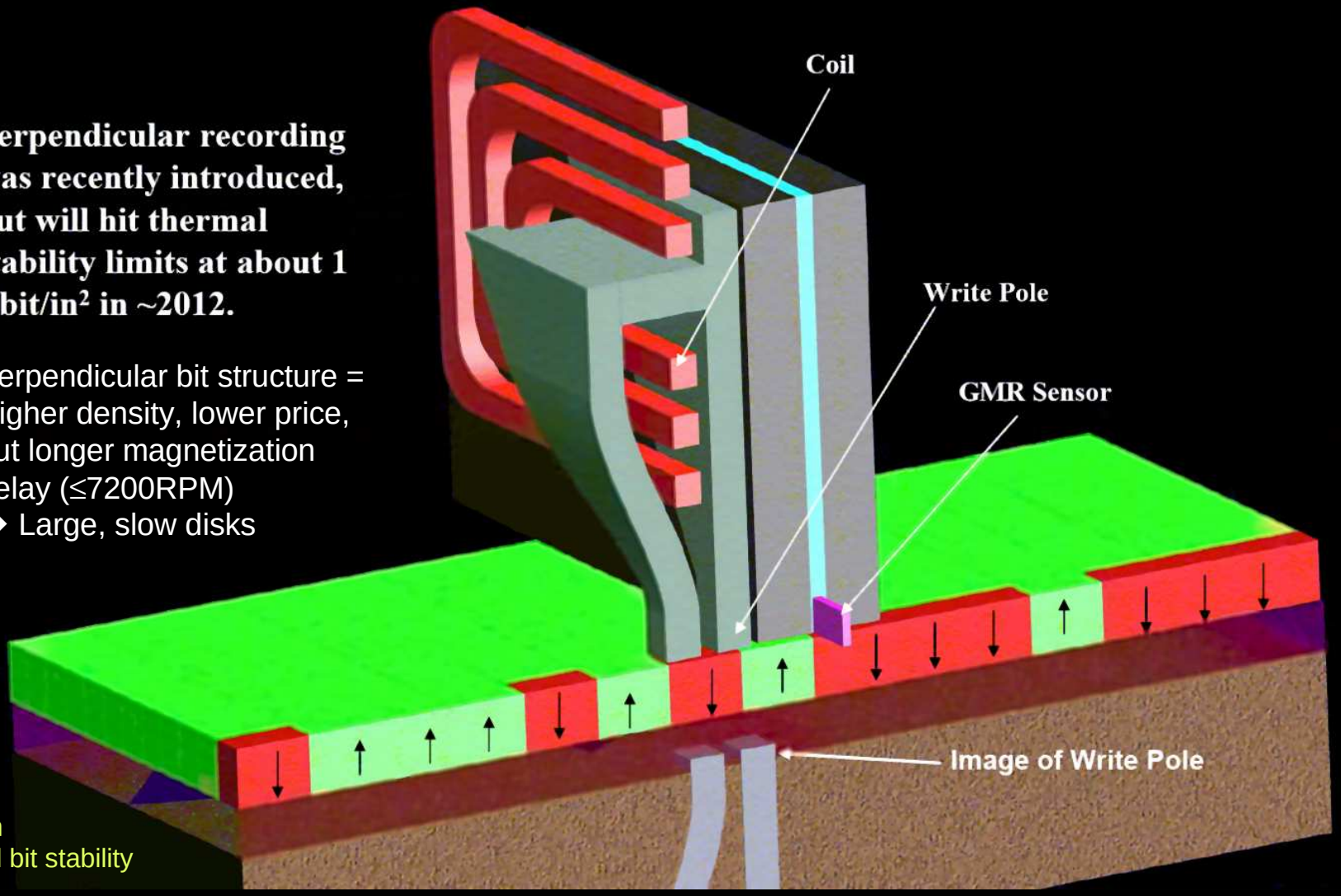
Storage in 2015

Decreasing Storage Technology Costs

Modern Disk Head for *perpendicular* Bits

Perpendicular recording was recently introduced, but will hit thermal stability limits at about 1 Tbit/in² in ~2012.

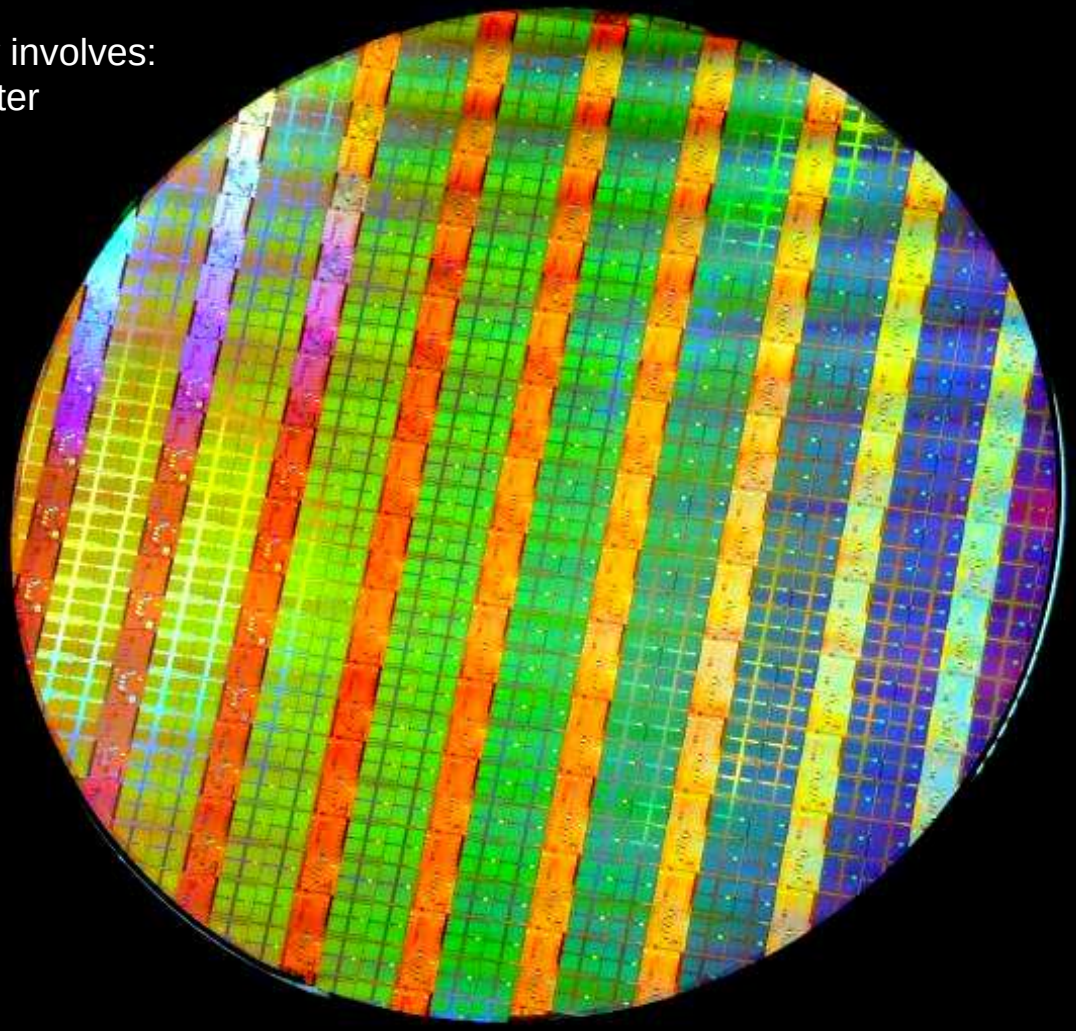
Perpendicular bit structure = Higher density, lower price, but longer magnetization delay (≤7200RPM)
➔ Large, slow disks



(!) Rotation
(!) Thermal bit stability

Modern Chip Wafer (30cm Ø, 45 nm Structure)

Cheaper chip memory involves:
1. Larger wafer diameter
2. Thinner lithography
3. More bits/transistor



(!) Diameter handling
(!) Lithography < 40nm
(!) Lower yield < 40nm

Image: Intel

Disk Storage vs. Chip Memory

Flash Memory prevails in mobile Systems

Disk

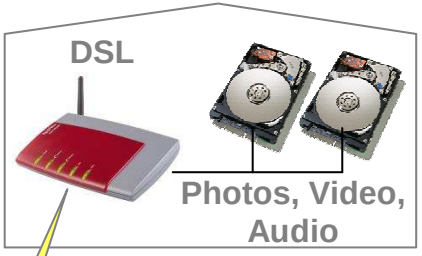
Flash

Image: Apple

Online Bandwidth will enable Flash PCs

with Disk:
Capacity

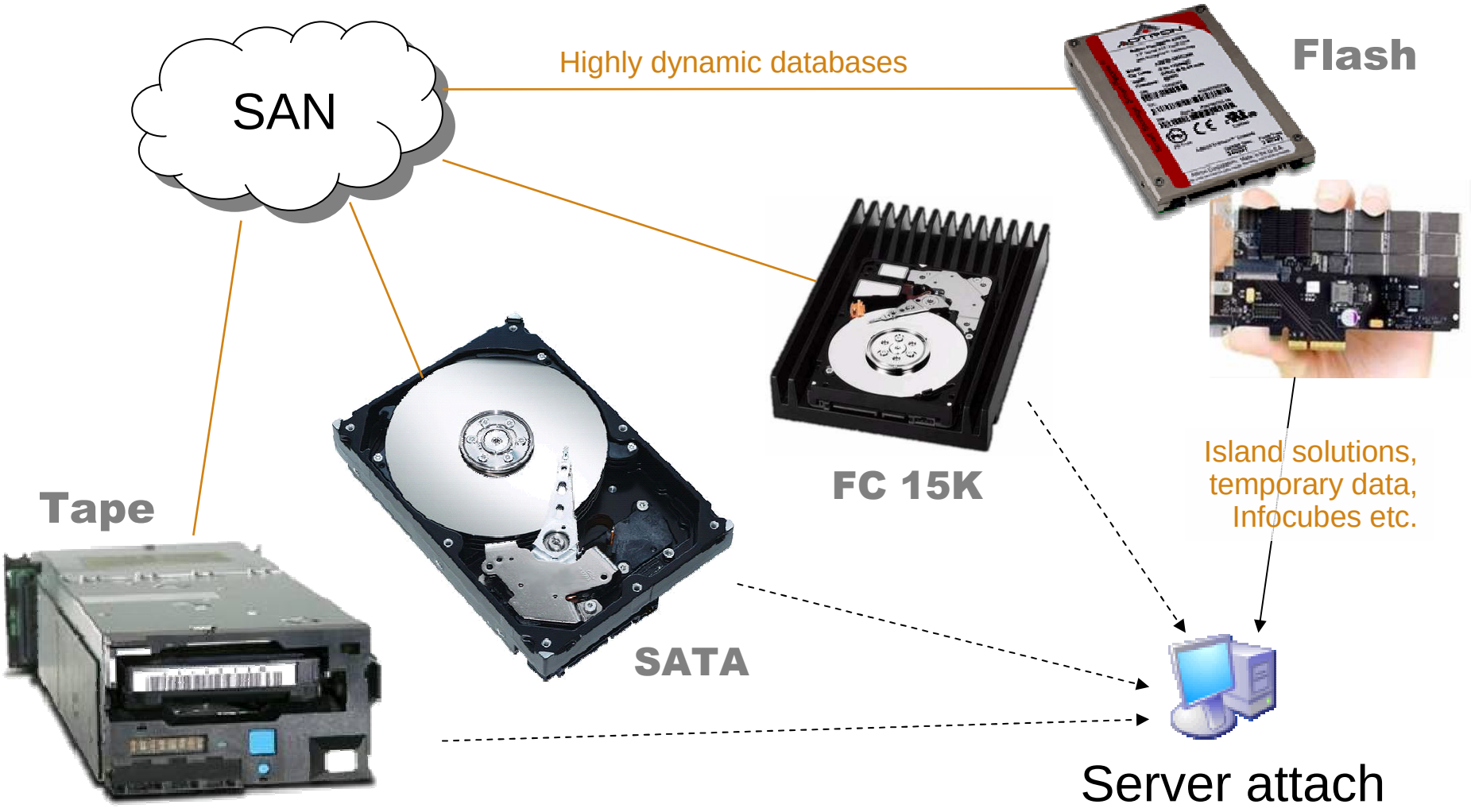
Photos, Video, Audio, Presentations, Documents, Software



with Flash:
Low Power

Go with less capacity →
Move surplus to network

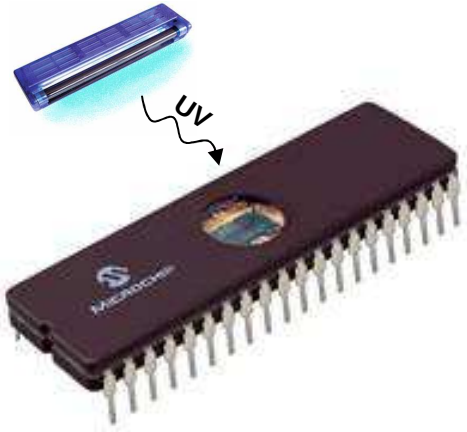
"Flash" Storage Tier in the Data Center



Chip Memory Technology Limits

Flash-Memory is built from EEPROM Blocks

EPROM
(1970)

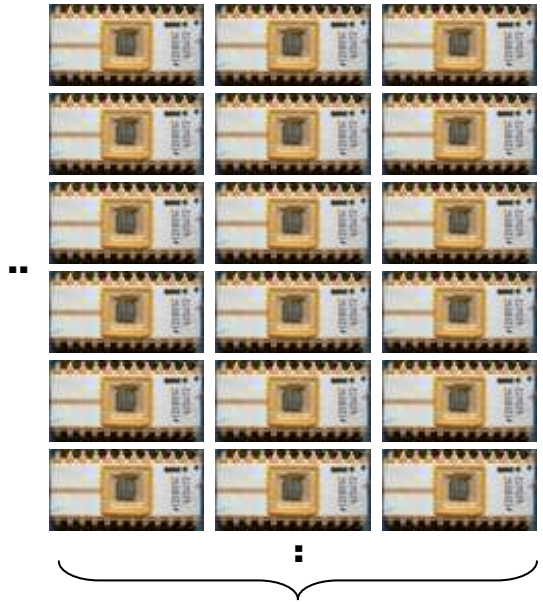


Delete with UV,
rewrite with 12 V

EEPROM

Delete with 12V,
rewrite with 5 V

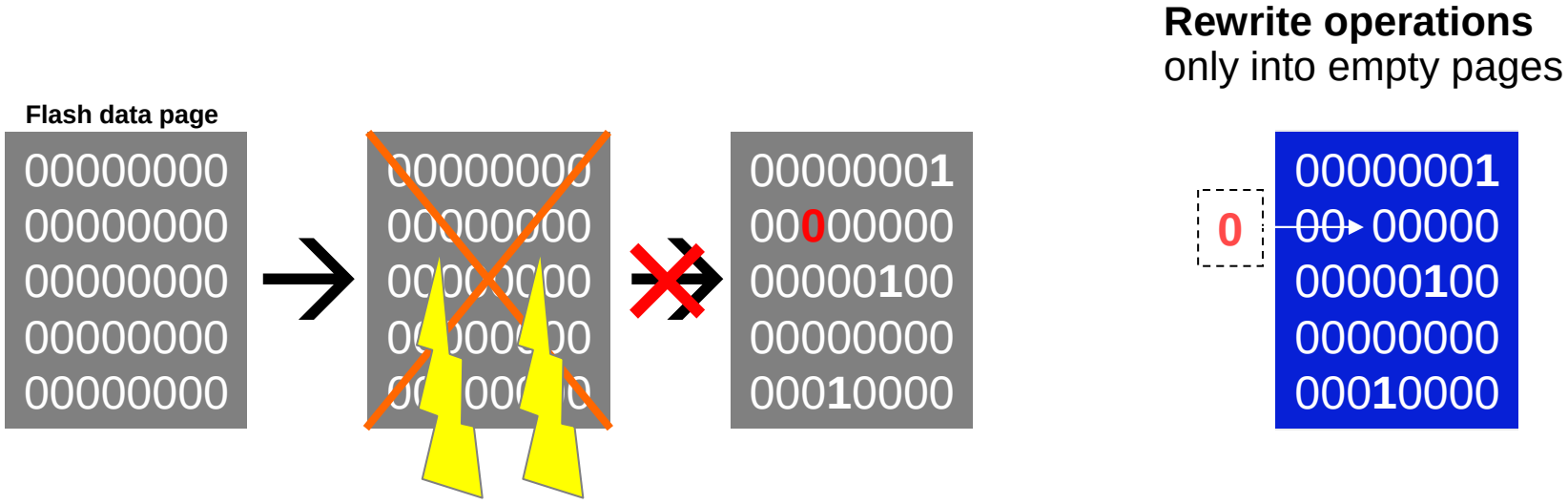
Flash Memory



Block-wise
deletion
with 10V

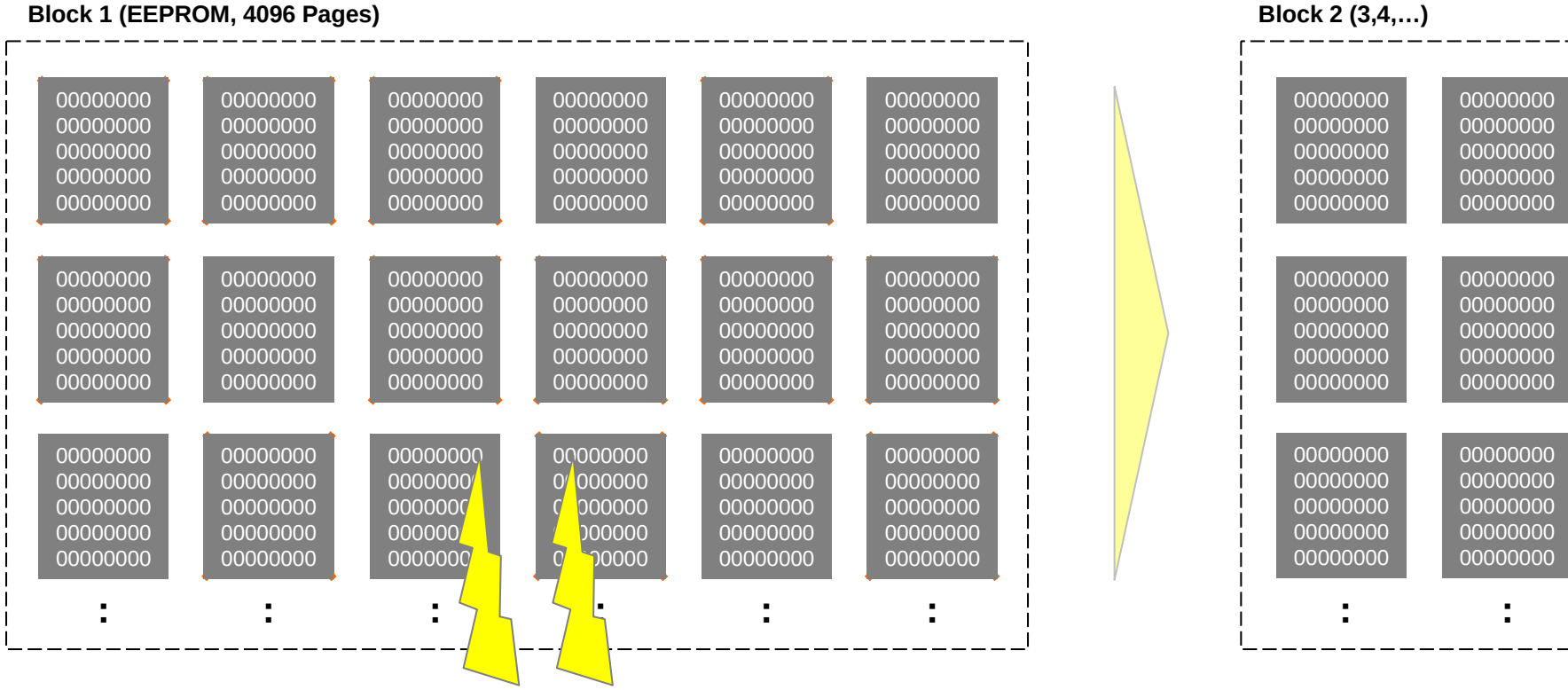
Flash Memory writes to empty Pages only

1 Deletion Pulse per Rewrite = Wear Stress



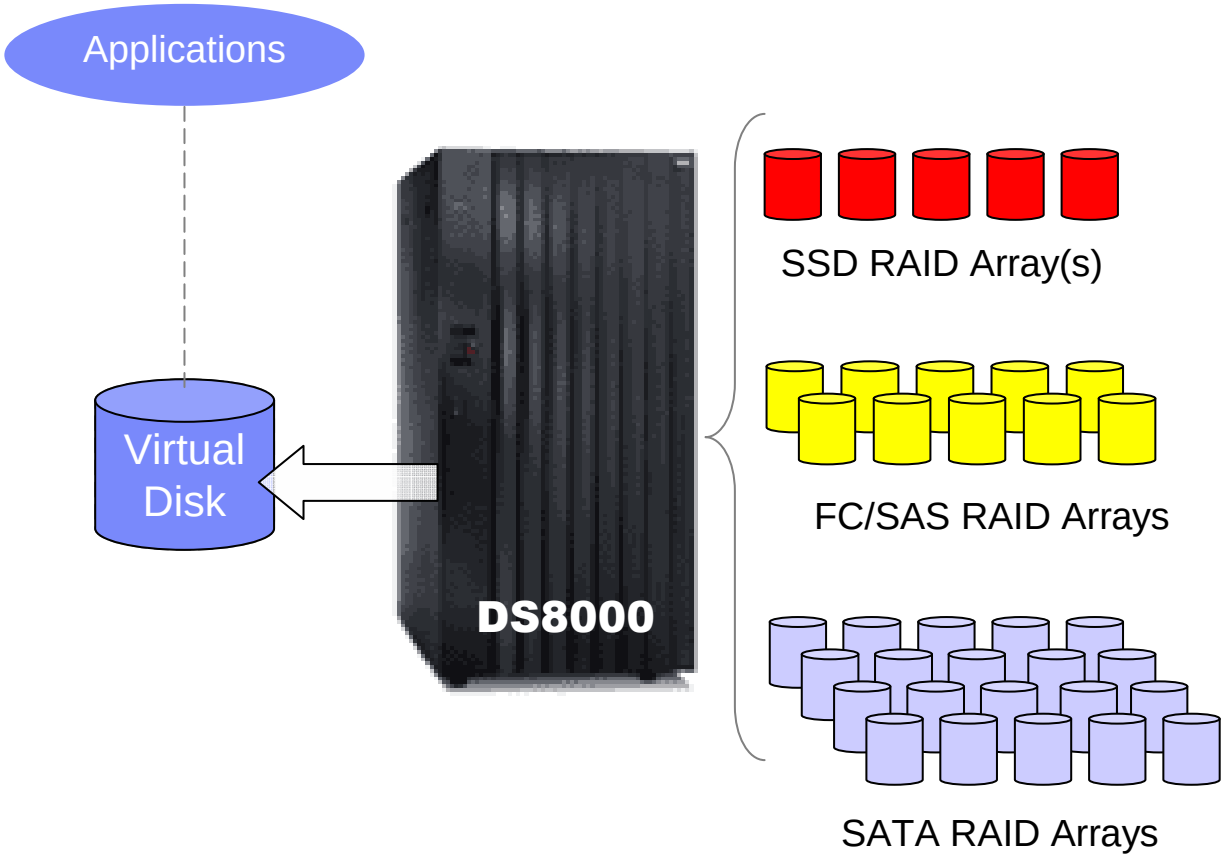
- Old used pages remain **locked**, awaiting deletion
- Max. 100.000 deletion pulses per page → distribute to minimize wear
- Optimal cache algorithms must target wear minimization

"Garbage Collection" → Fair Wear Levelling



- Fewer erasure cycles → longer life & better performance
- Wear levelling for over-utilized and under-utilized cells
- **Significant over-provisioning → higher price**

Use Flash Memory wisely !

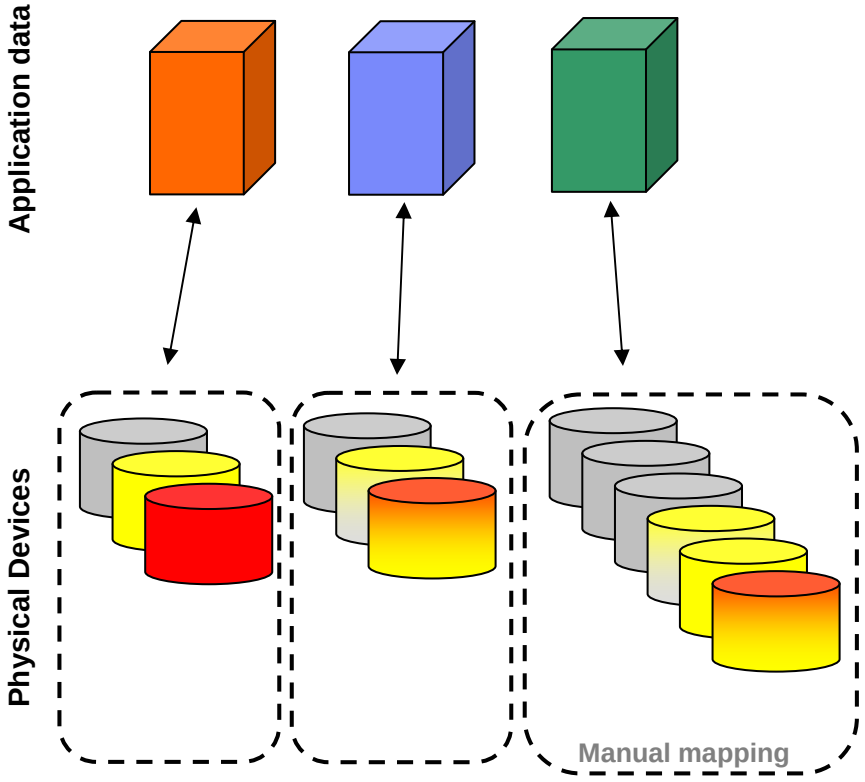


- Lowest Watt / IOPS
  & footprint / IOPS
- Lowest € / IOPS
- Lowest € / GB

Optimal Efficiency, less administrative Effort

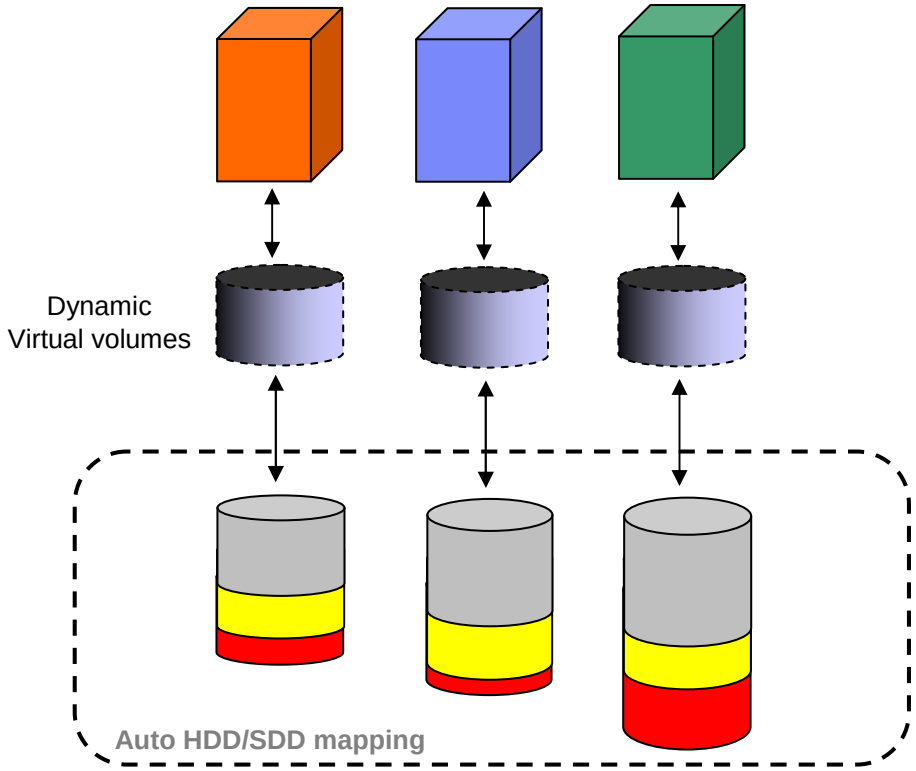
Preview

Traditional disk mapping



Volumes have different characteristics. Applications need to place them on correct tiers of storage based on usage.

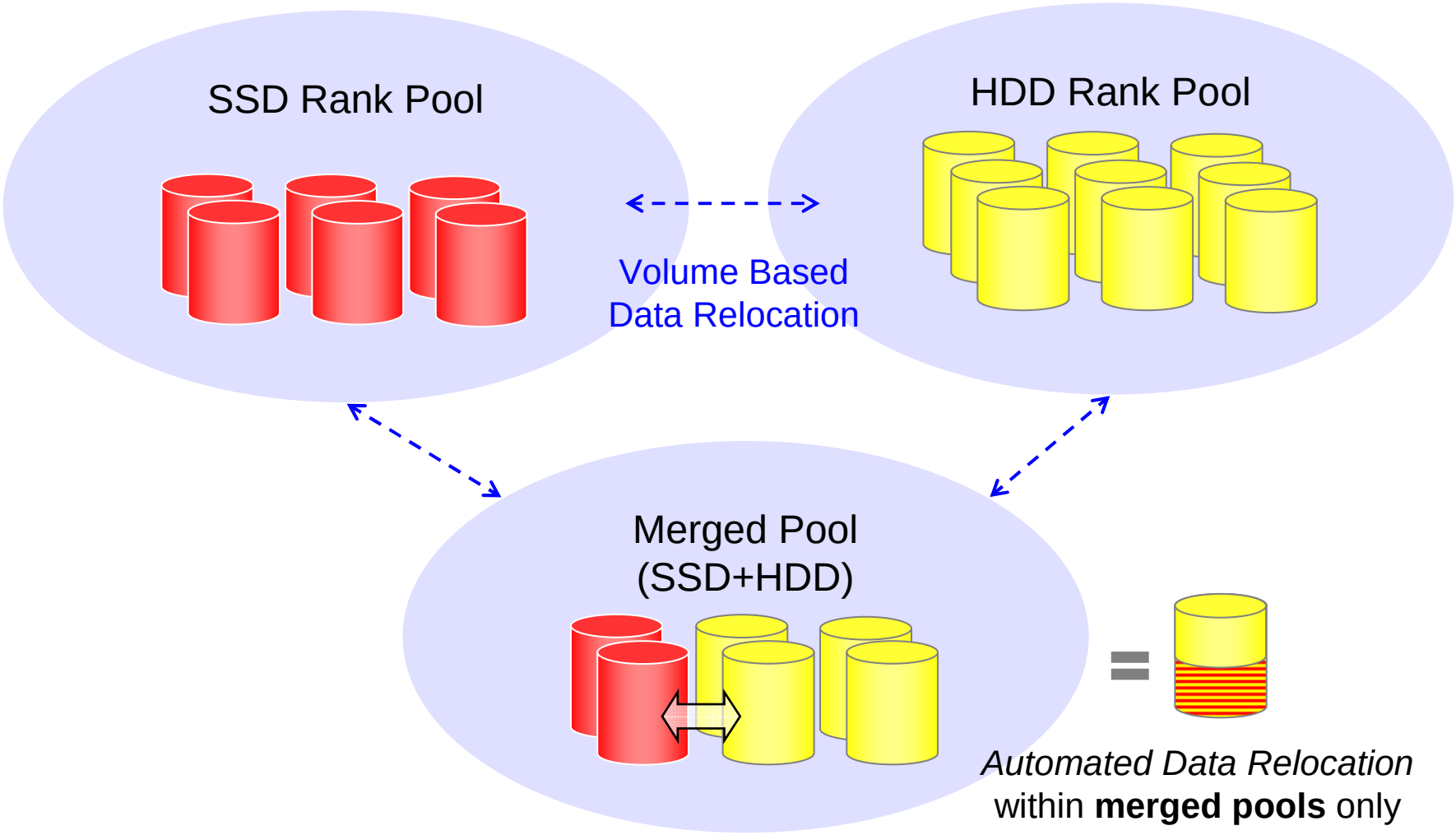
Dynamic Virtual Storage mapping



All volumes appear to be “logically” homogenous to apps. But data is placed at the right tier of storage based on its usage through smart data placement and migration.

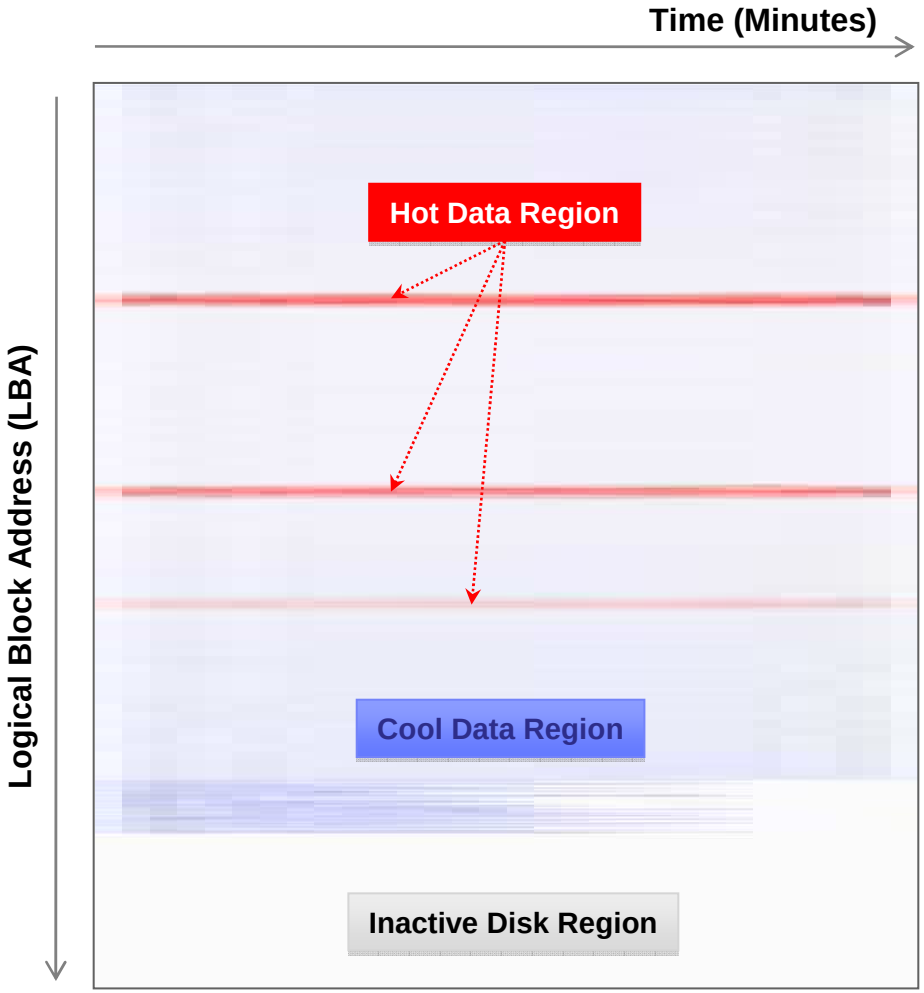
Volume- and Extent-based Relocation

Preview

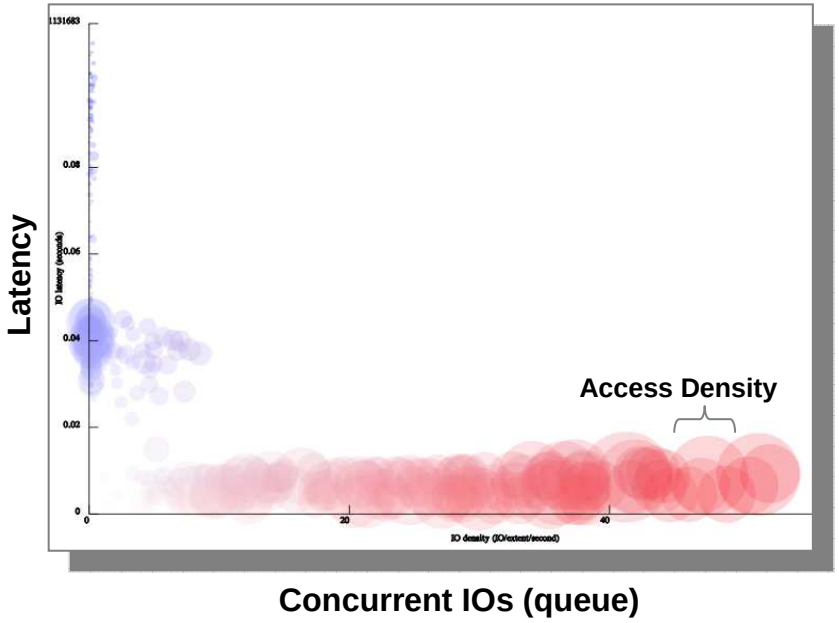


Preview

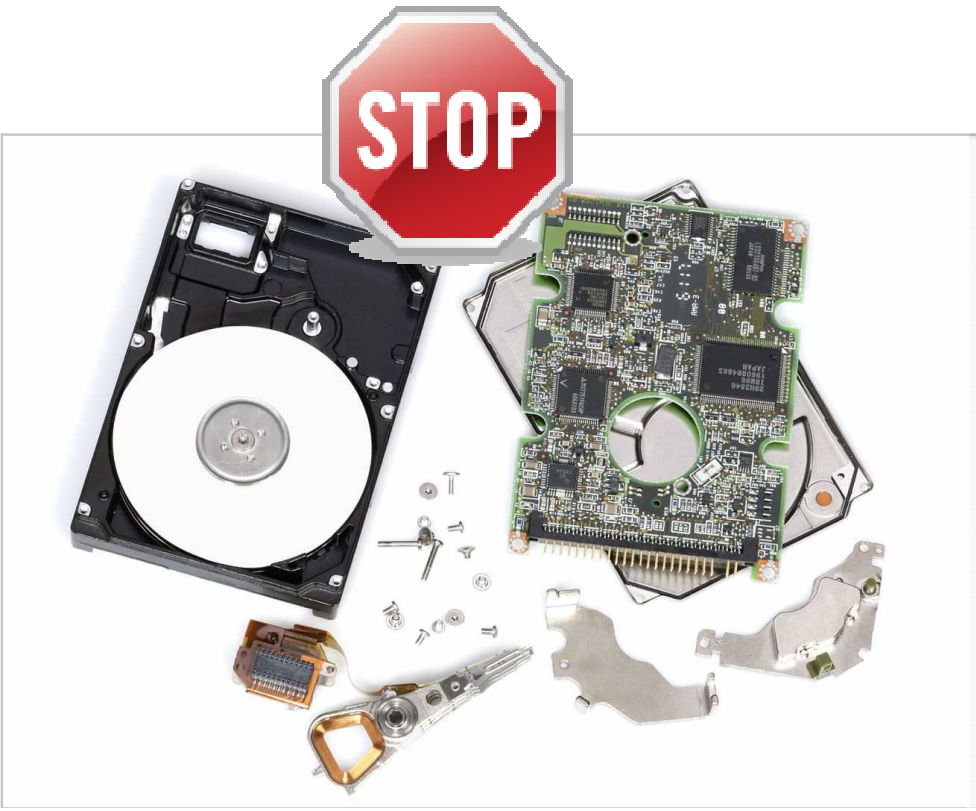
Optimize for highest IOPS or lowest Latency?

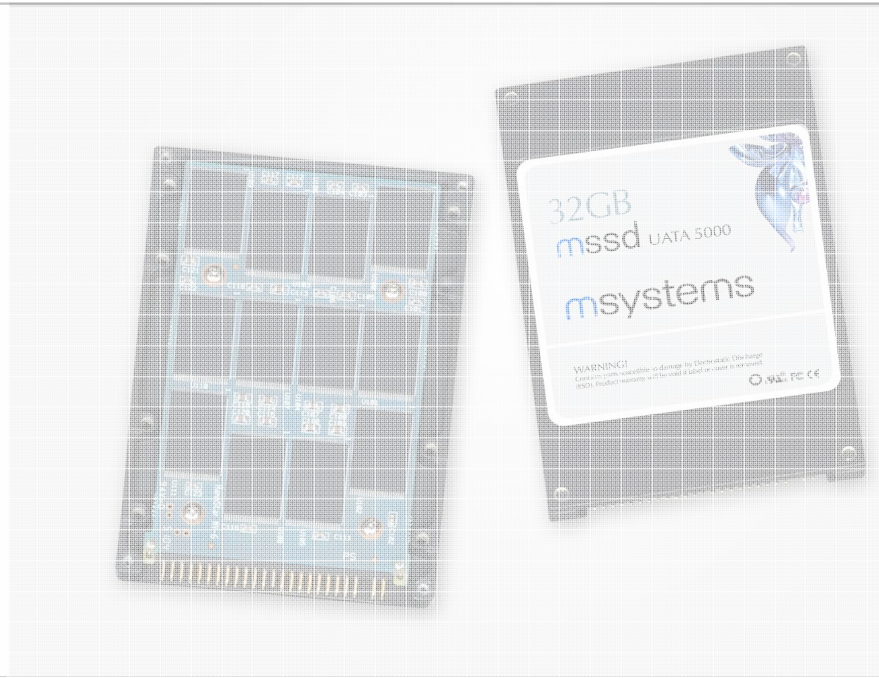


- Fine grained average performance monitoring
- Workload learning
- Hotspot auto-elimination

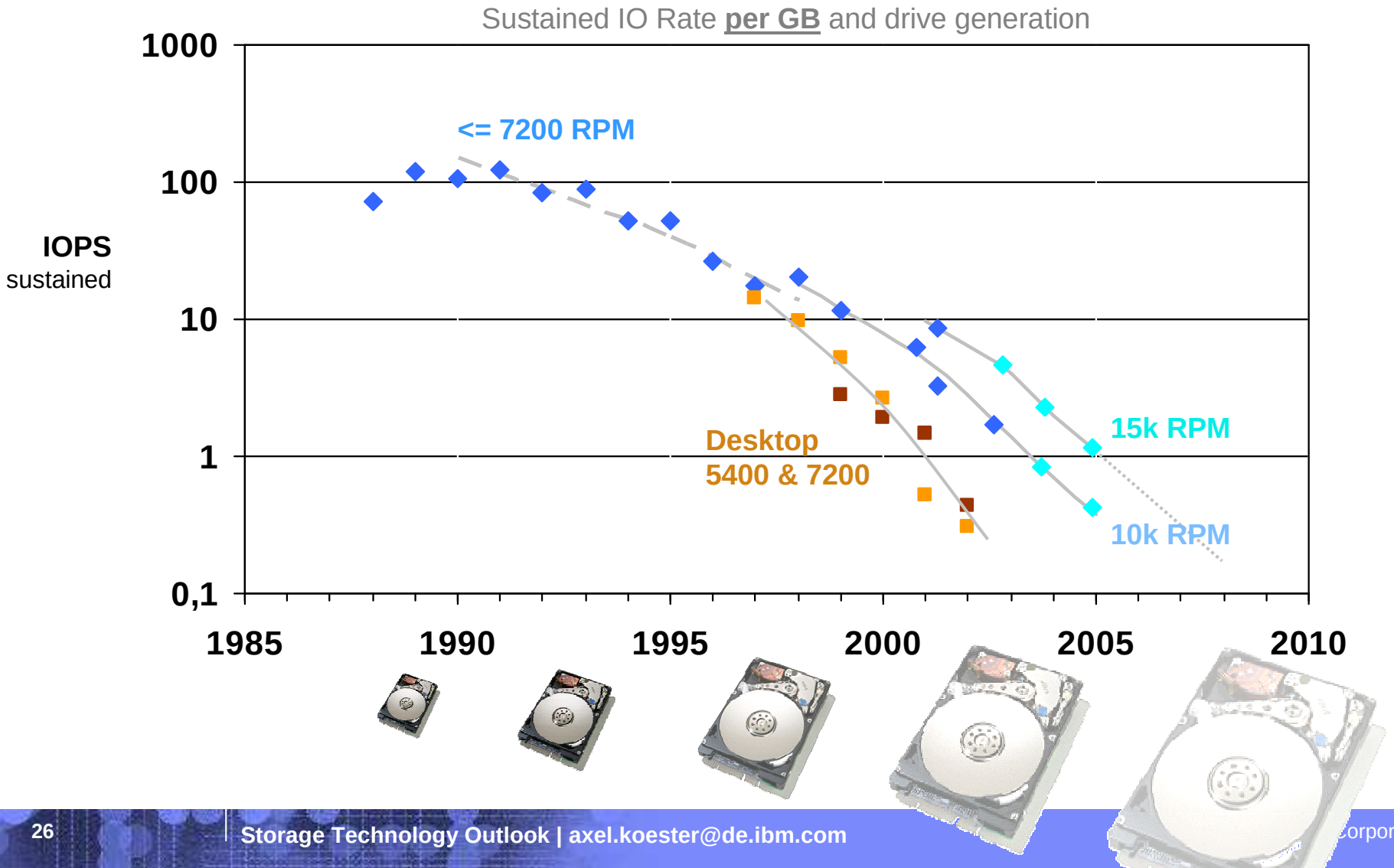


Disk Technology Limits

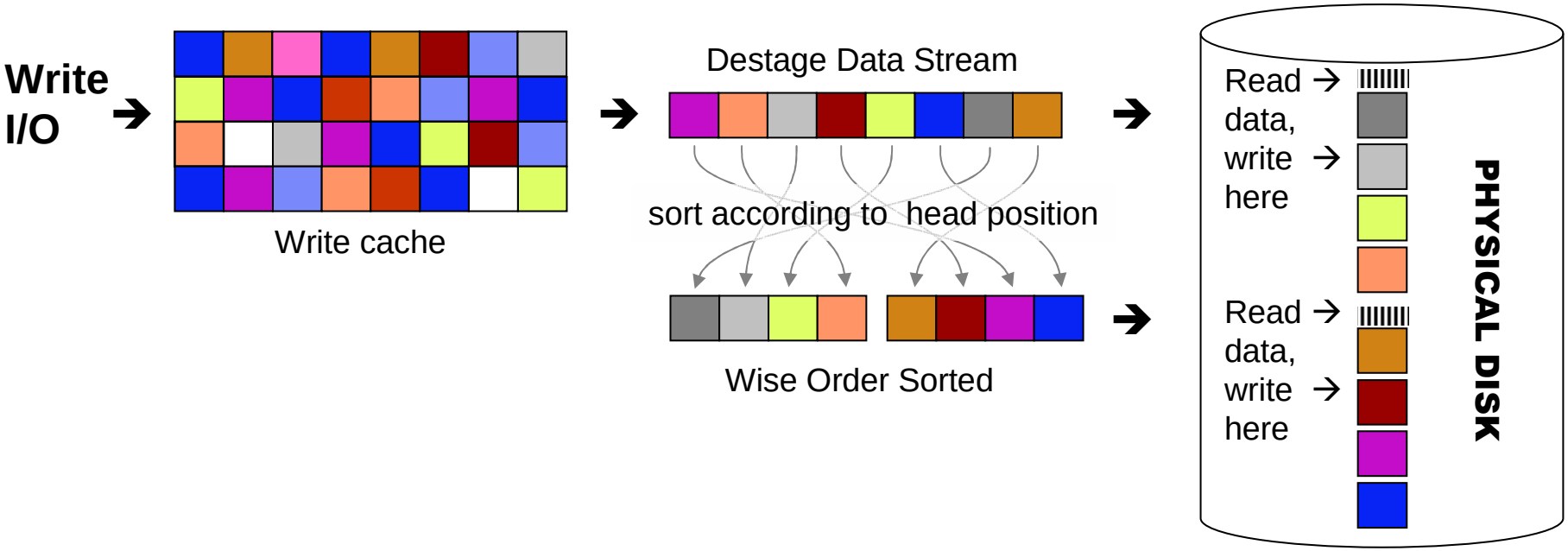




Harddisk Data Rate *per Gigabyte* drops alarmingly



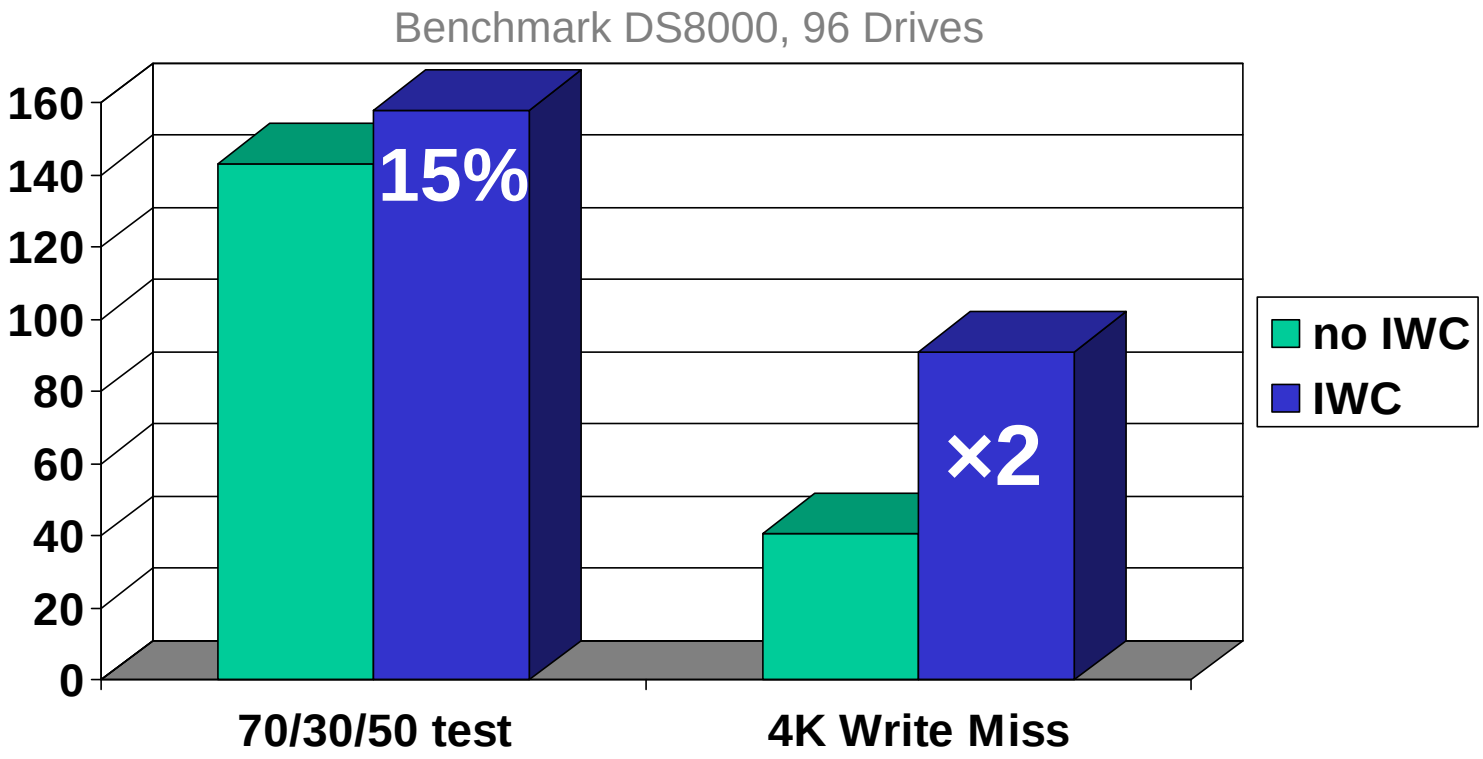
New: Intelligent Write Caching *(since DS8000 R4.2)*



- Optimized for **minimal head movement**
- Delay writes in cache until head is in proximity
- Ideal for **random databases**

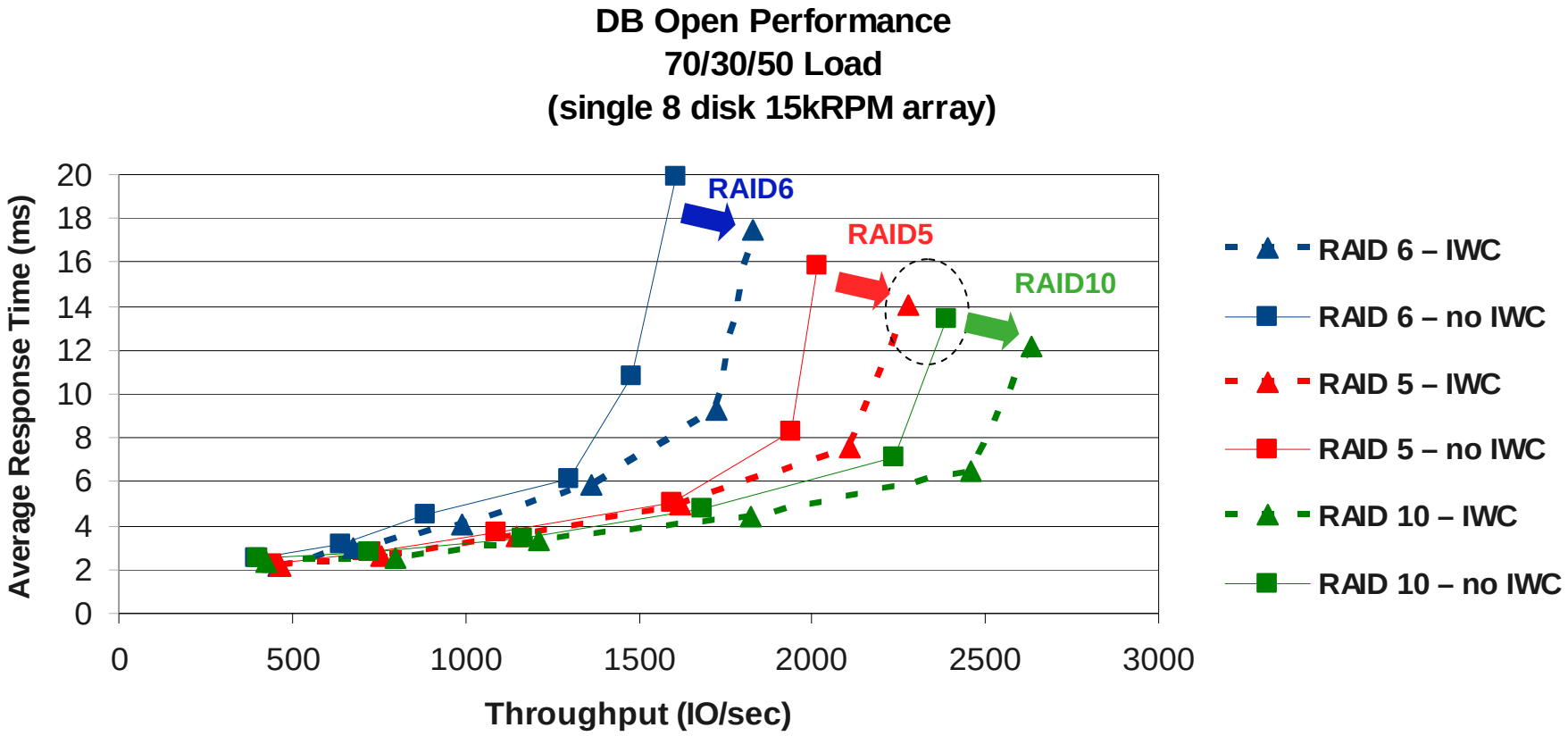
"Wise Order Writes"

Wise Order Writes – for Databases



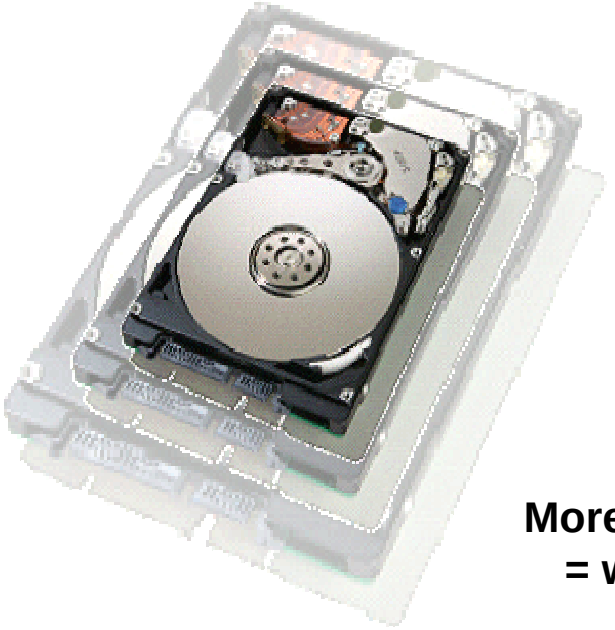
RAID 10, 15K RPM, mix of 146, 300 and 450 GB
(64) (16) (16)

Wise Order Writes – new RAID 5 at RAID10 Speed



RAID 5 for DB workload is now almost as fast as "old" RAID 10.

Thin Provisioning = counter-productive ?



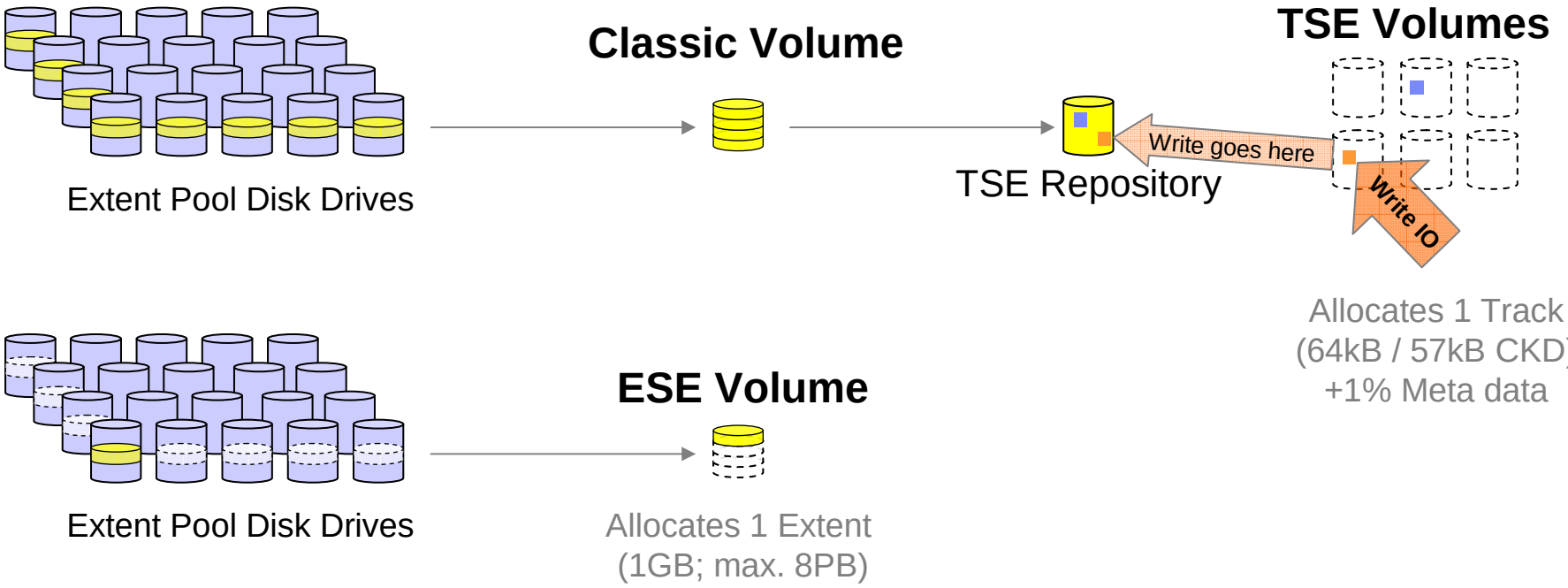
More virtual GB / drive
= worse IOPS / GB

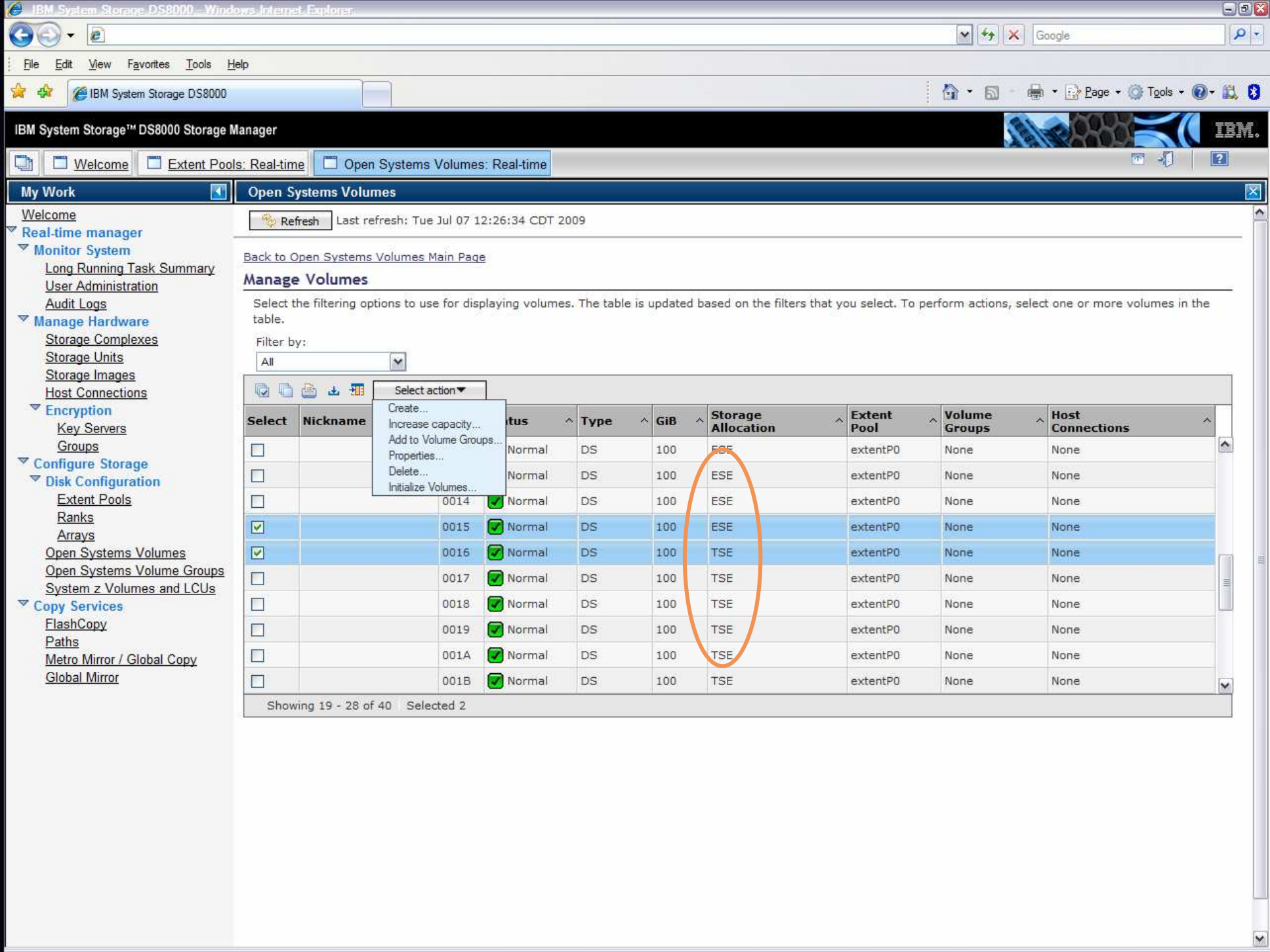
DS8000 Thin Provisioning : Extent or Track Level ?

Extent Space Efficiency uses existing extent pool virtualization.
Extents are allocated on first write only, without host wait impact.

Track Space Efficient is optimized for FlashCopy SE & Global Mirror.

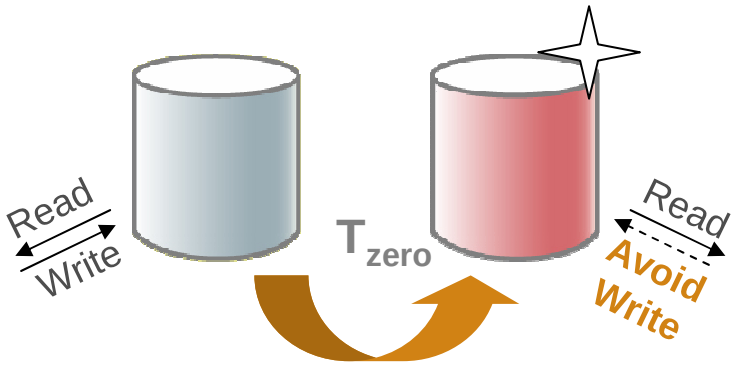
Use TSE for volumes not expected to fill up to 100%, e.g. Flashcopy NOCOPY target volumes.



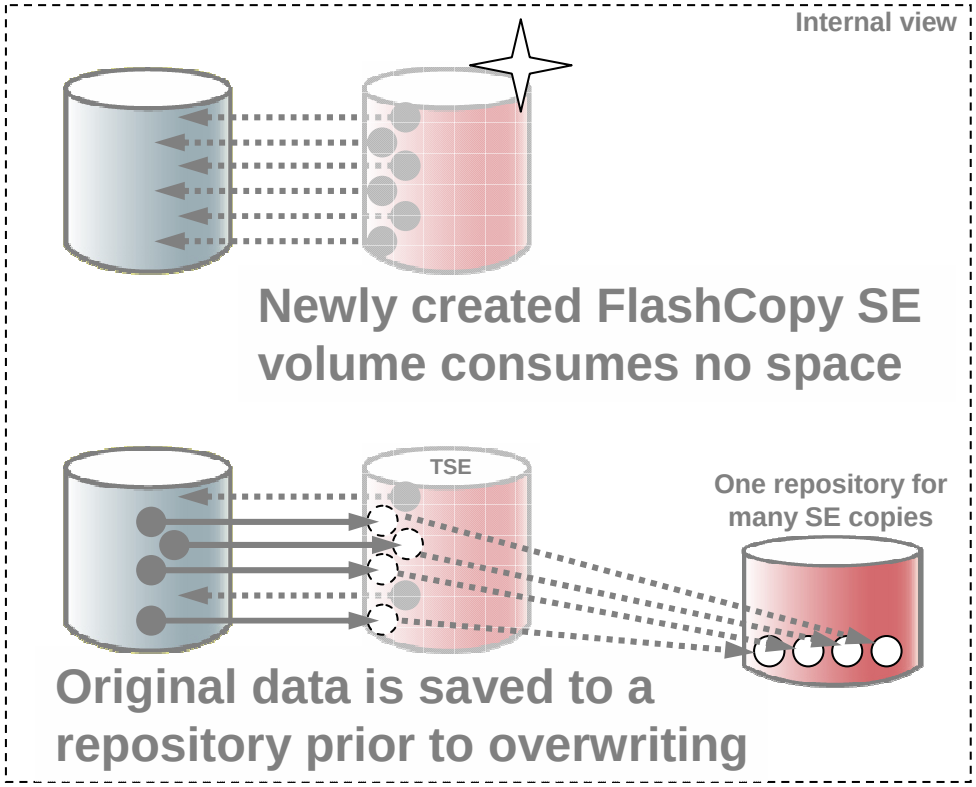

IBM System Storage DS8000 - Windows Internet Explorer
File Edit View Favorites Tools Help
IBM System Storage DS8000
IBM System Storage™ DS8000 Storage Manager
Welcome
Extent Pools: Real-time
Open Systems Volumes: Real-time
My Work
Welcome
Real-time manager
Monitor System
Long Running Task Summary
User Administration
Audit Logs
Manage Hardware
Storage Complexes
Storage Units
Storage Images
Host Connections
Encryption
Key Servers
Groups
Configure Storage
Disk Configuration
Extent Pools
Ranks
Arrays
Open Systems Volumes
Open Systems Volume Groups
System z Volumes and LCUs
Copy Services
FlashCopy
Paths
Metro Mirror / Global Copy
Global Mirror
Open Systems Volumes
Refresh
Last refresh: Tue Jul 07 12:26:34 CDT 2009
Back to Open Systems Volumes Main Page
Manage Volumes
Select the filtering options to use for displaying volumes. The table is updated based on the filters that you select. To perform actions, select one or more volumes in the table.
Filter by:
All
Select action
Create...
Increase capacity...
Add to Volume Groups...
Properties...
Delete...
Initialize Volumes...
Select
Nickname
tus
Type
GiB
Storage Allocation
Extent Pool
Volume Groups
Host Connections
Normal DS 100 ESE extentP0 None None
Normal DS 100 ESE extentP0 None None
0014 Normal DS 100 ESE extentP0 None None
0015 Normal DS 100 ESE extentP0 None None
0016 Normal DS 100 TSE extentP0 None None
0017 Normal DS 100 TSE extentP0 None None
0018 Normal DS 100 TSE extentP0 None None
0019 Normal DS 100 TSE extentP0 None None
001A Normal DS 100 TSE extentP0 None None
001B Normal DS 100 TSE extentP0 None None
Showing 19 - 28 of 40 Selected 2

FlashCopy SE = "Flashcopy onto a TSE Volume"

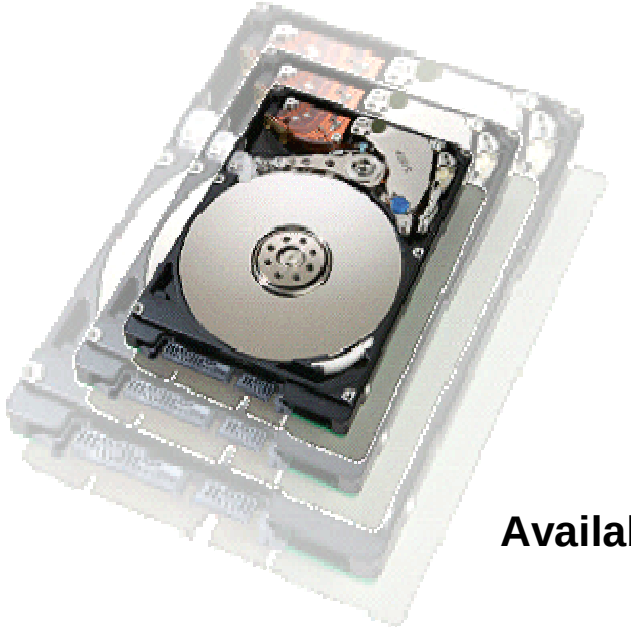
FlashCopy SE uses a common repository (e.g. per extent pool) to save original data when resolving references. Grain = 64kB; 57kB CKD; +1% Metadata



Instant Space-Efficient FlashCopy

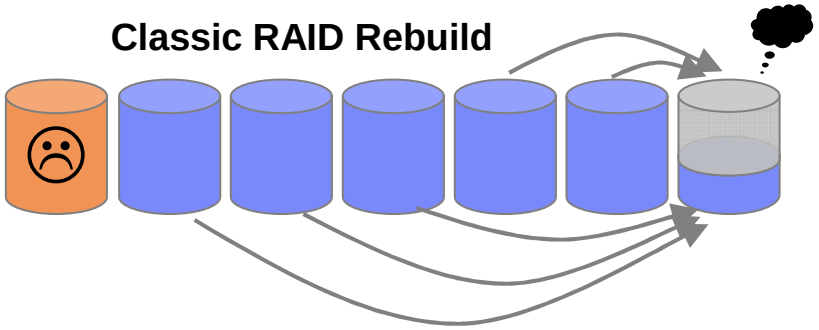


Better System Availability despite larger Disks

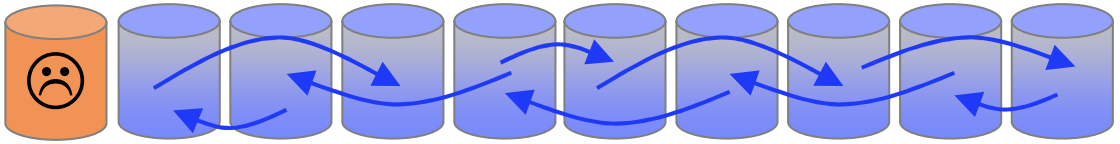


Availability = f(drive speed)

The Challenge: Terabyte Hot Spare Drive Rebuild



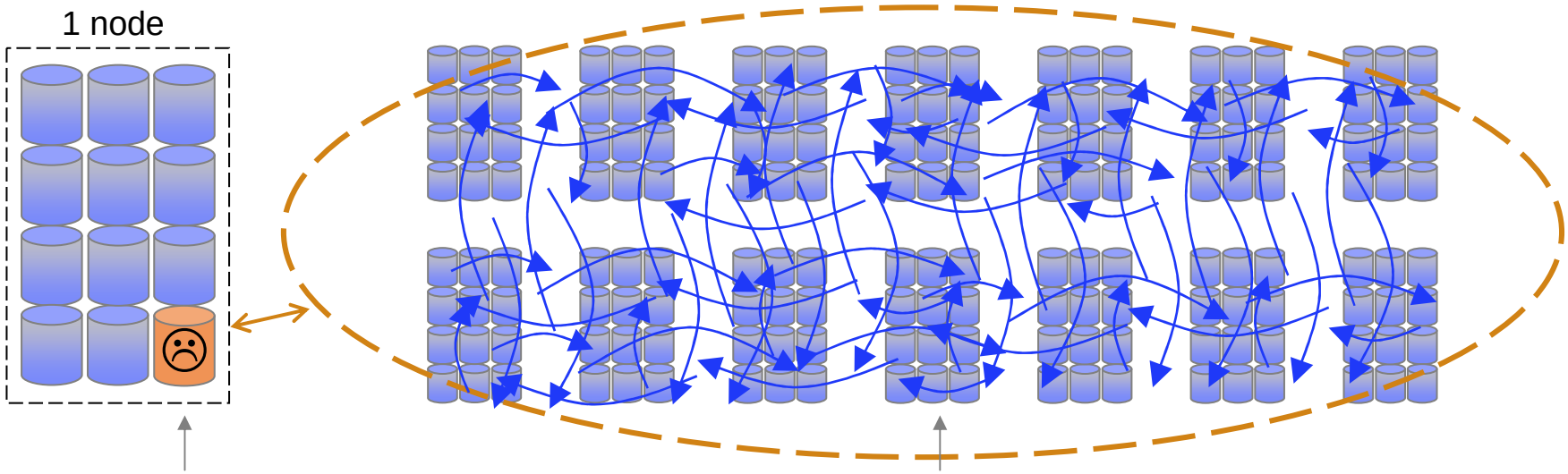
The solution: **Declustered RAID** – parallel recovery of redundancy



Each disk handles 1/10th of the rebuild load = 1/10th of the time

Even less with more disks involved

The XIV Redundancy Principle



The redundancy data of any disk is distributed across all other disks not in the same node

Rebuilding lost redundancy is a non-impacting activity of all disks

Rebuilding redundancy works many times in-a-row. Disk redundancy and node redundancy use the same mechanism. Worst case rebuild time for 1TB ≤ 30min.

Storage in 2015: Non-volatile RAM

Image © DOCB

The Suburb Commuter Problem

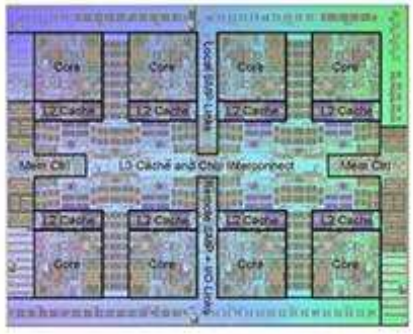
Expensive real estate, but short paths

Cheap real estate, but traffic delays

€ **CPU cycles**

L1 cache: 1 **DRAM: 100** **Flash: 10^5-10^6** **Disk: 10^7-10^8**

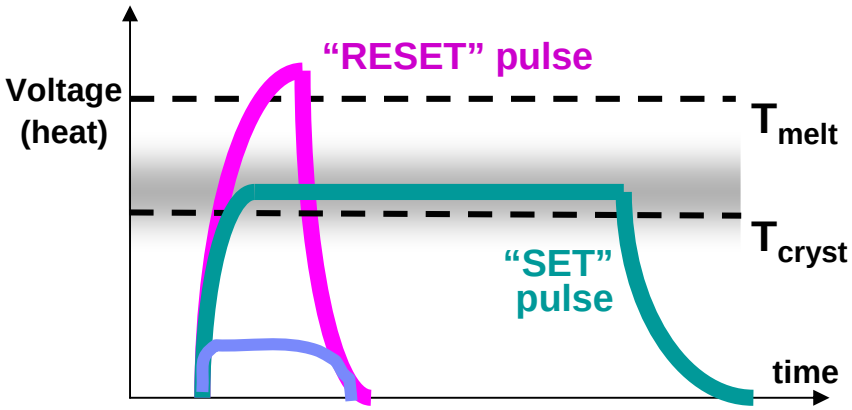
L2/L3 cache: 10-50

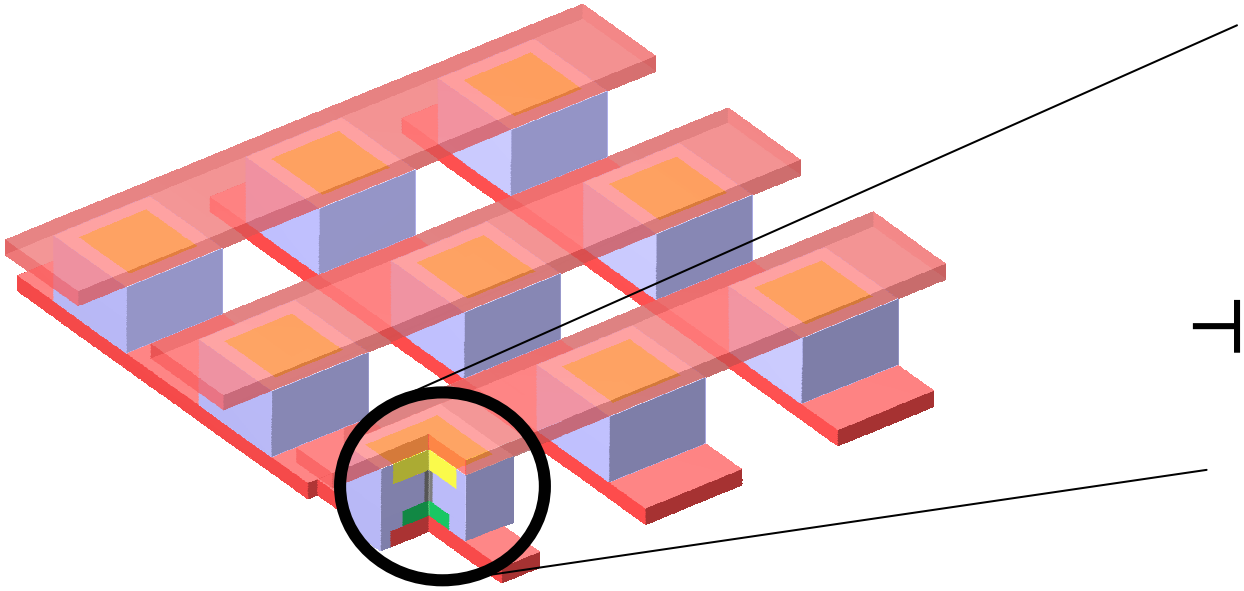


GAP

Non-volatile storage

Non-volatile RAM Candidate: Phase-change RAM



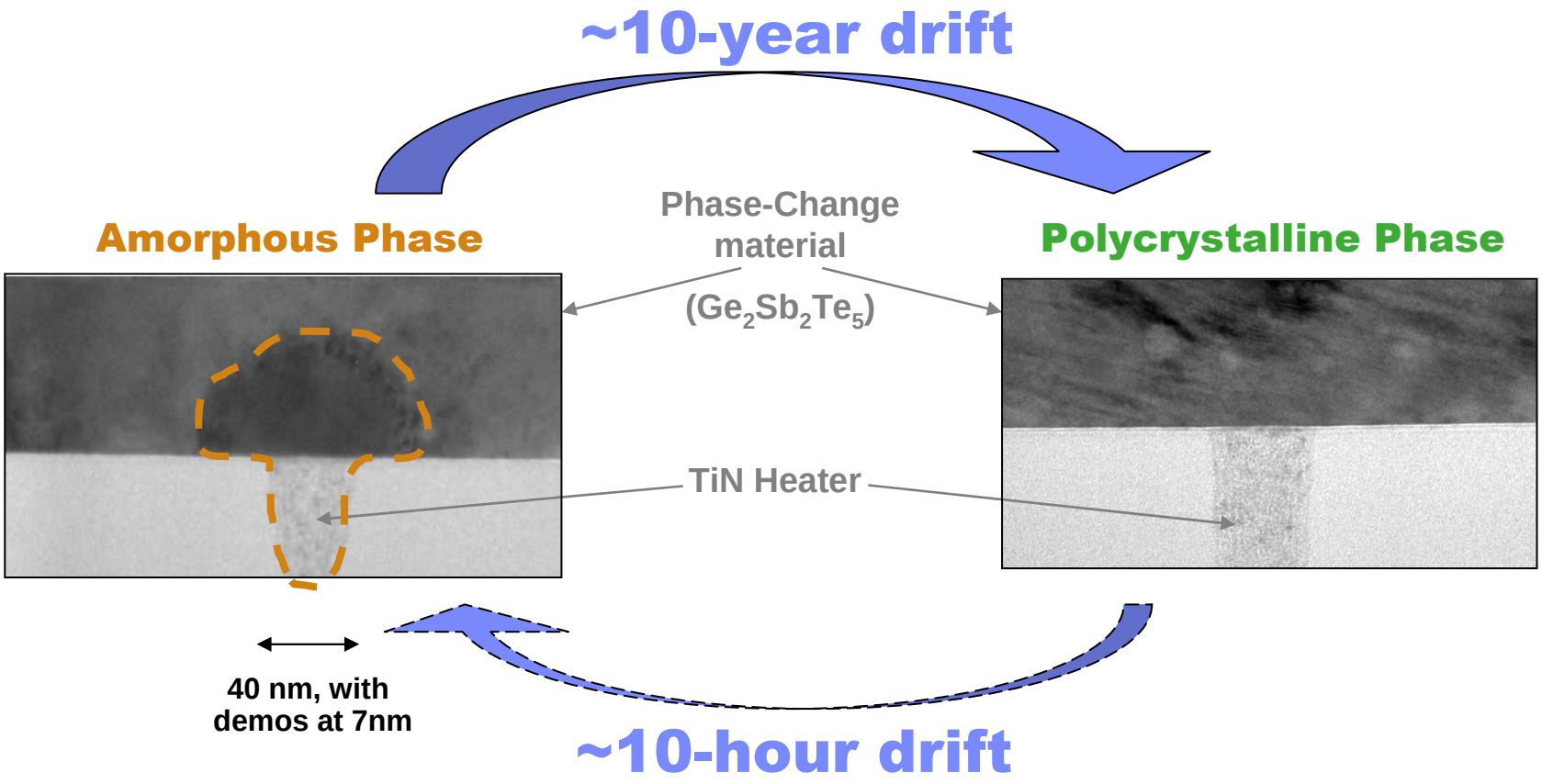


PCRAM
"programmable resistor"

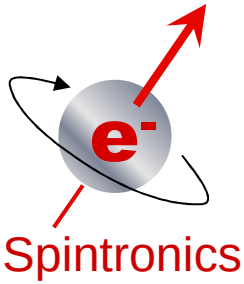
Access device
(Transistor)

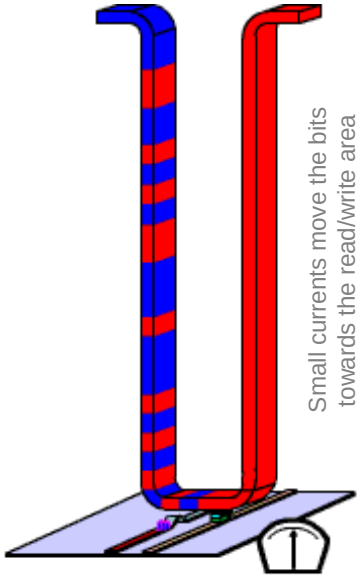
Currently sets the scaling limit ($4F^2$)

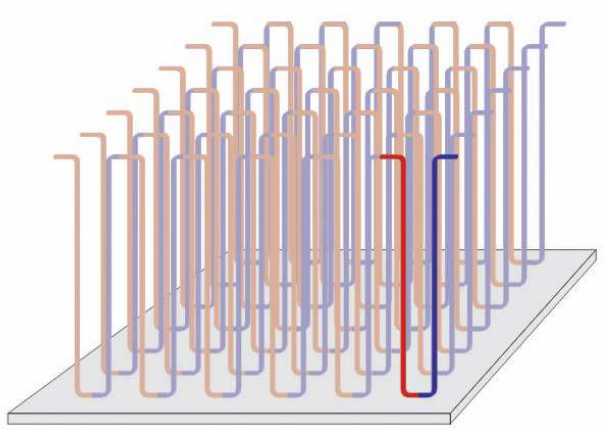
Phase-change "Nano Mushroom"



Spintronics : "RaceTrack" Memory



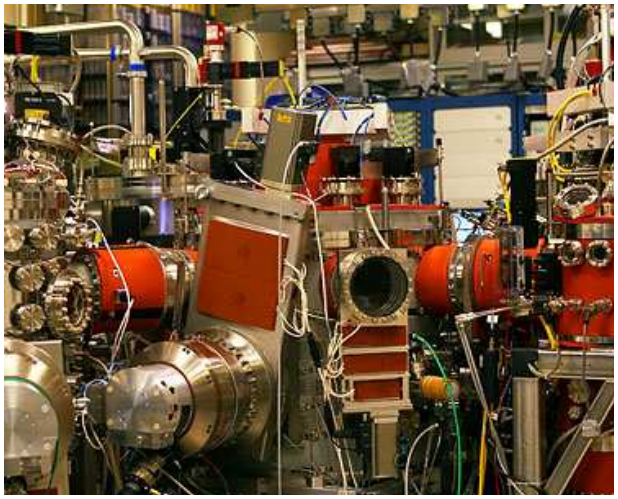




RaceTrack Storage Array

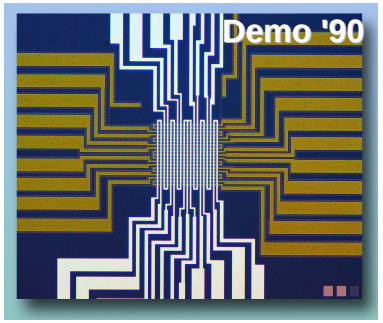
Storage in 3rd Dimension

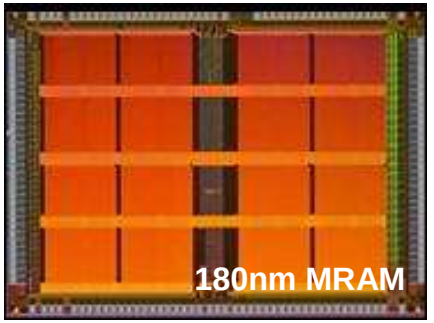
"Large" read/write head, "small" bits on ferroelectric nano wire

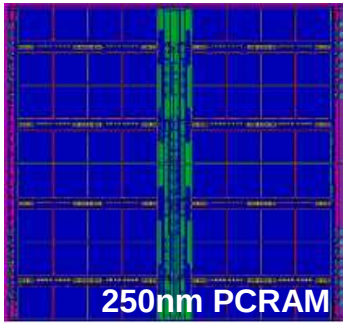


IBM Fellow Stuart Parkin, Inventor of GMR read heads, investigates "Racetrack Memory"

Non-volatile RAM : Many Players, many Options



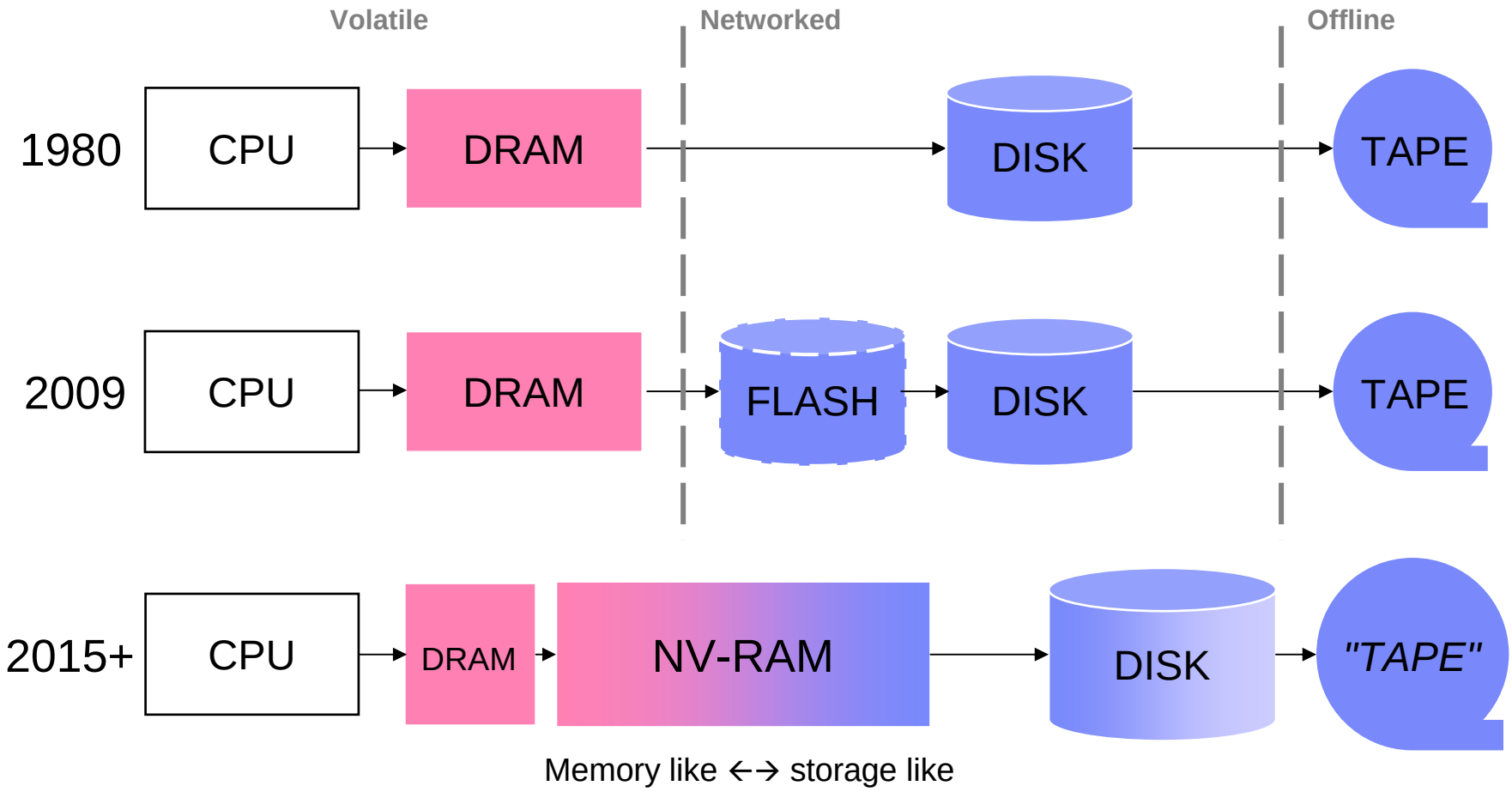




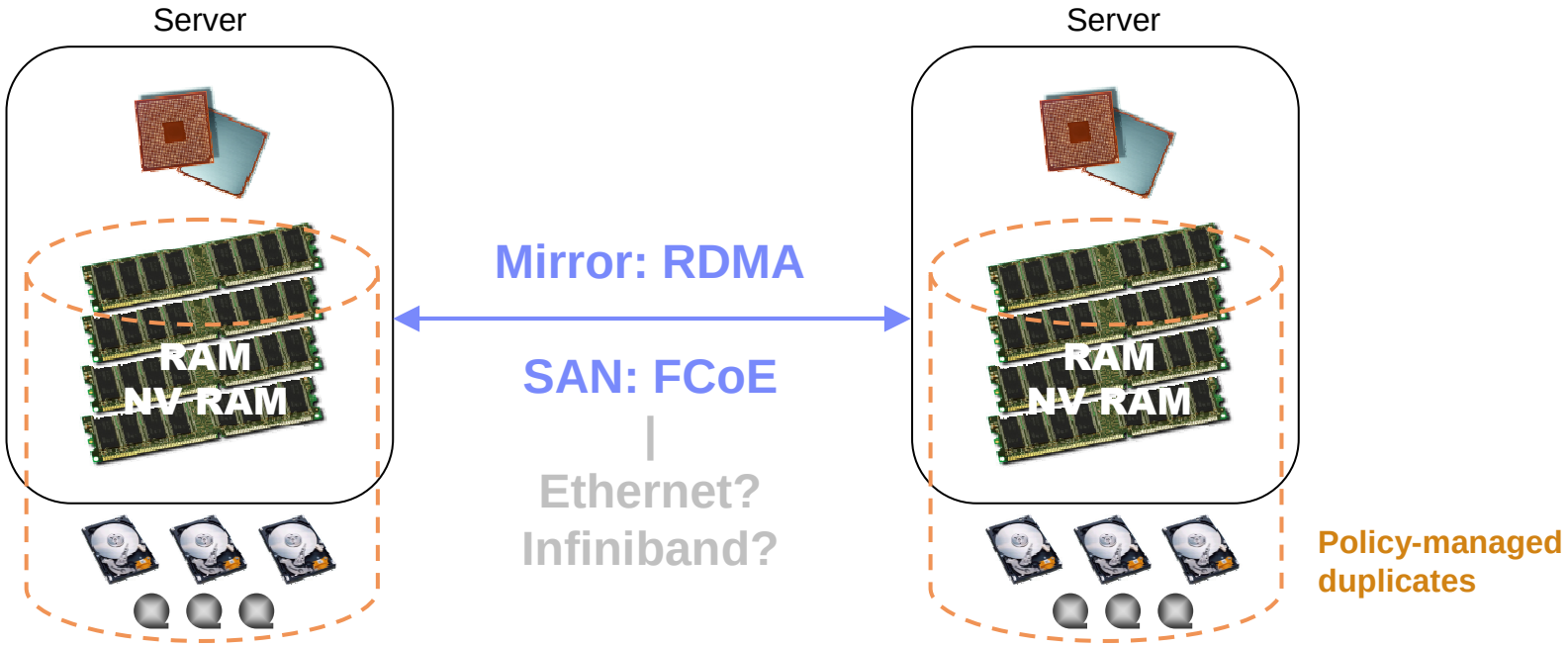
No wear, high reliability, high speed

- Phase-Change RAM
- Magnetic RAM
- Ferroelectric RAM
- Nano-mechanical RAM
- …

Storage will become an integral Part of the Server



Networked Non-Volatile RAM in Serve/Store Combo



- The *SAN* of the future needs to provide connectivity to peer system's Non-Volatile RAM
- Experience and technology do exit today in high-performance computing (RDMA = Remote Direct Memory Access)

axel.koester@de.ibm.com