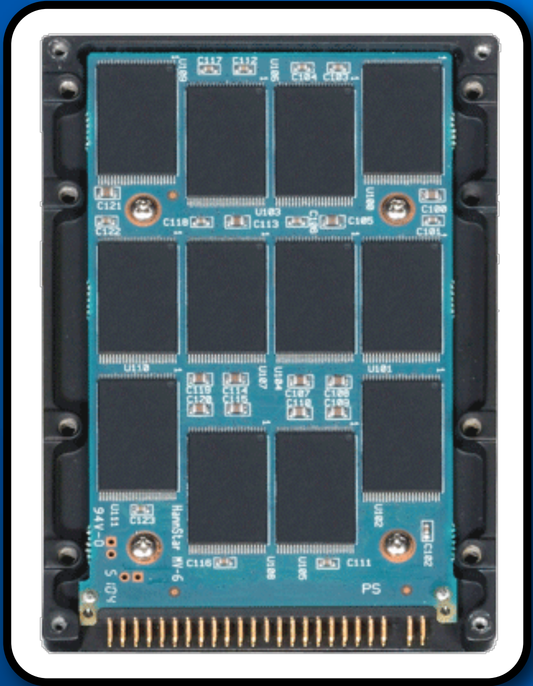
Seminar Energy-Efficient Databases
29.06.2011

Enhancing Energy Efficiency of Database Applications Using SSDs

Felix Martin Schuhknecht

Motivation

vs.



Motivation

Why Solid State Disks (SSDs) ?

Motivation

Why Solid State Disks (SSDs) ?

Positive:

- High Perfomance
- Low Power Consumption

Motivation

Why Solid State Disks (SSDs) ?

Positive:
- High Perfomance
- Low Power Consumption

Negative:
- Small Capacity
- Very Expensive

Motivation

Why Solid State Disks (SSDs) ?

Positive:

- High Perfomance
- Low Power Consumption

Negative:

- ~~Small Capacity~~
- Very Expensive

Motivation

Why Solid State Disks (SSDs) ?

Positive:

- High Perfomance
- Low Power Consumption

Negative:

- ~~Small Capacity~~
- ~~Very~~ Expensive

Motivation

Why Solid State Disks (SSDs) ?

Really?

Positive:

- High Perfomance
- Low Power Consumption

Negative:

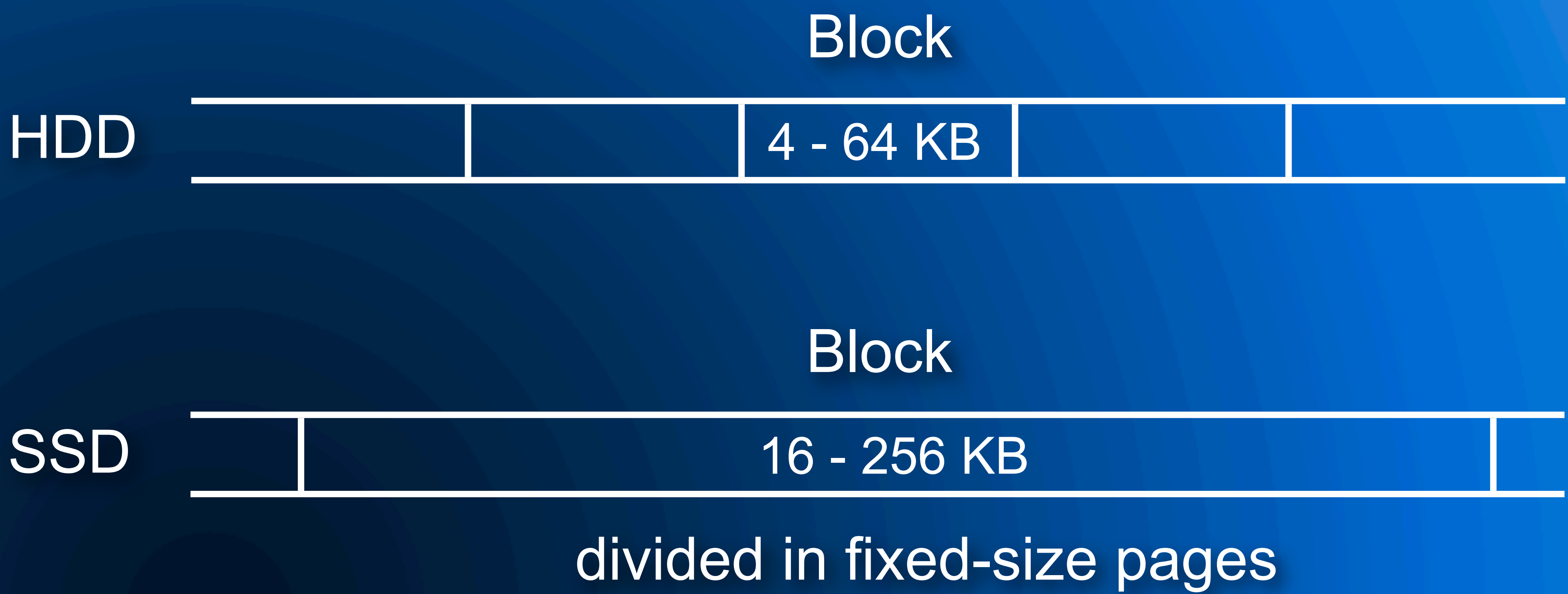
- ~~Small Capacity~~
- ~~Very~~ Expensive

Architecture: Flash versus Disk

	HDD	SSD
Cache	YES	NO
Limited Write Endurance	NO	YES
Block Size	SMALL	LARGE

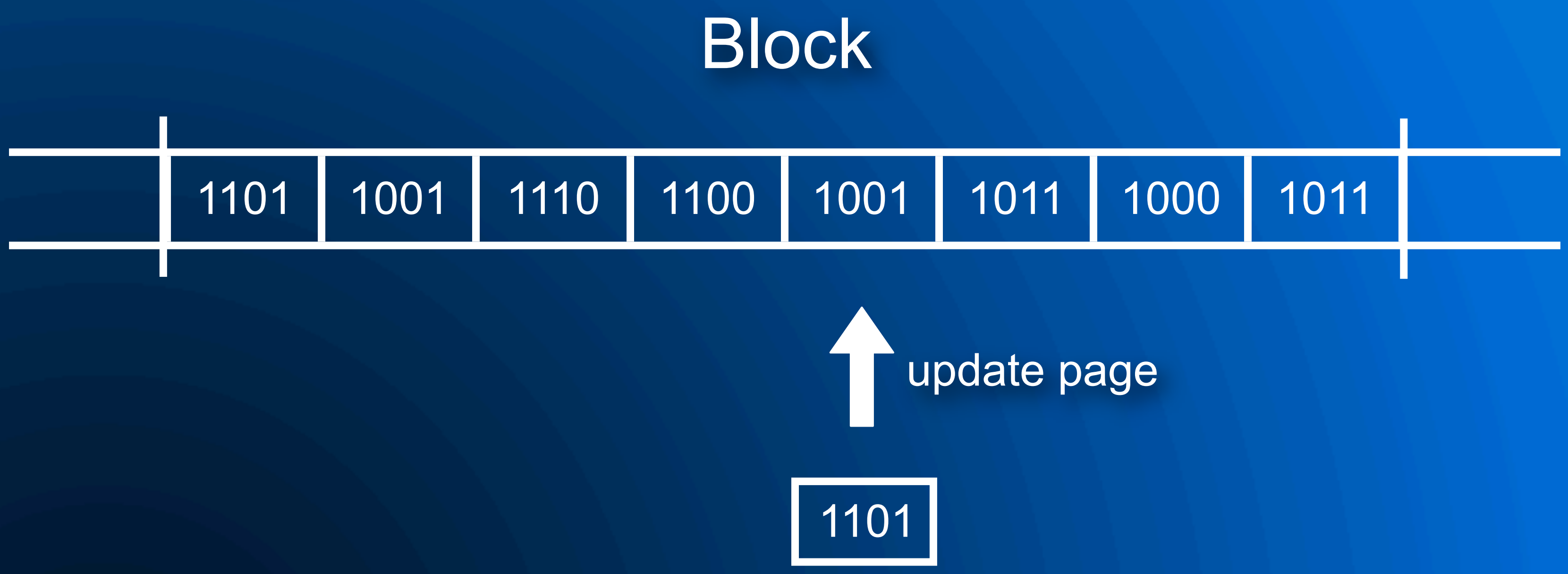
Architecture: Flash versus Disk

Block Size:



Architecture: Flash versus Disk

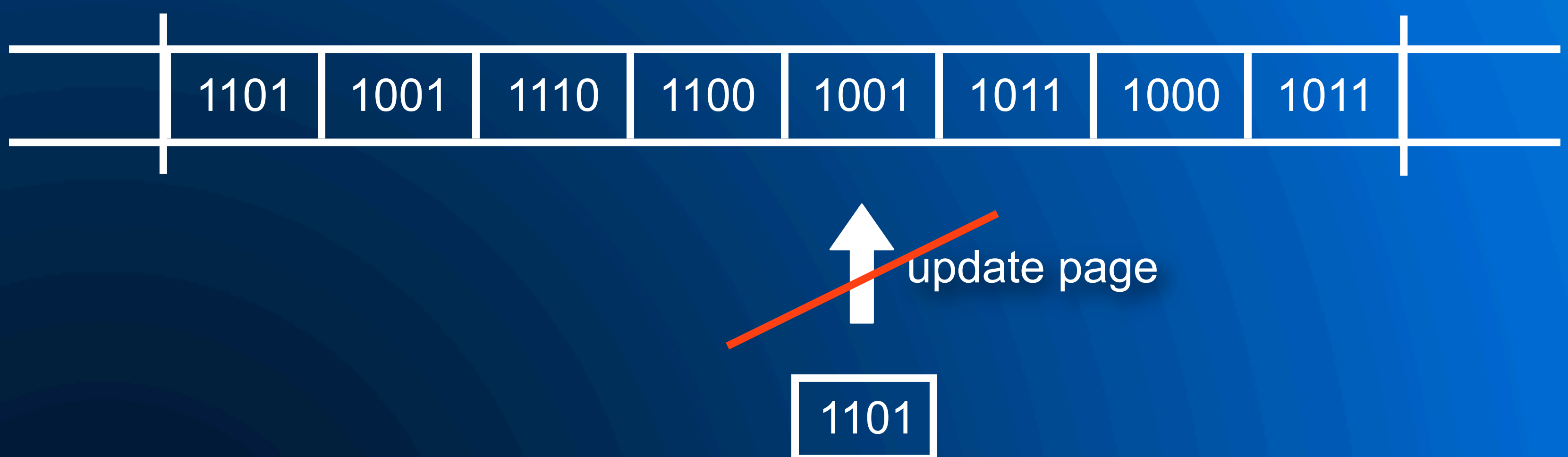
Problem: Page Updates



Architecture: Flash versus Disk

Problem: Page Updates

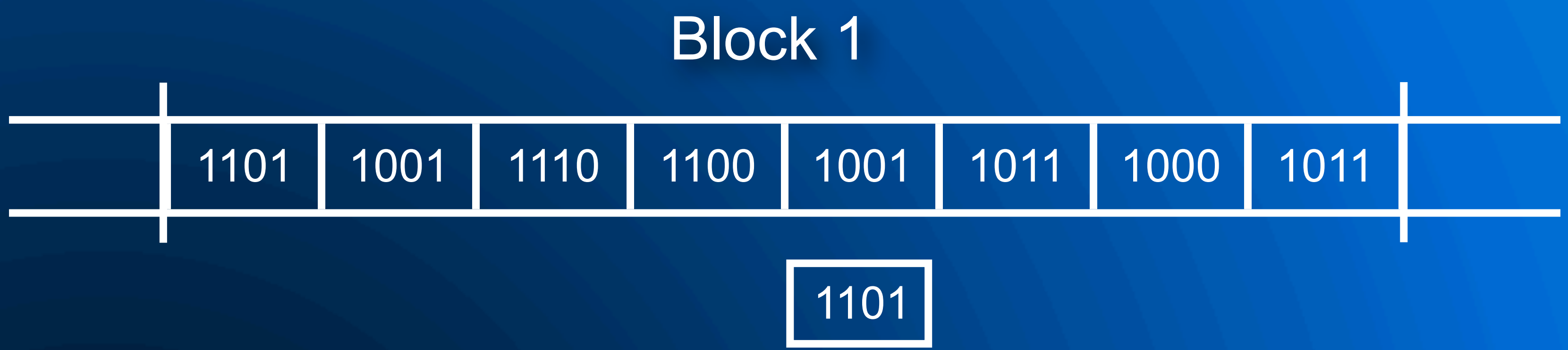
Block



Only transitions from 1 to 0 possible!

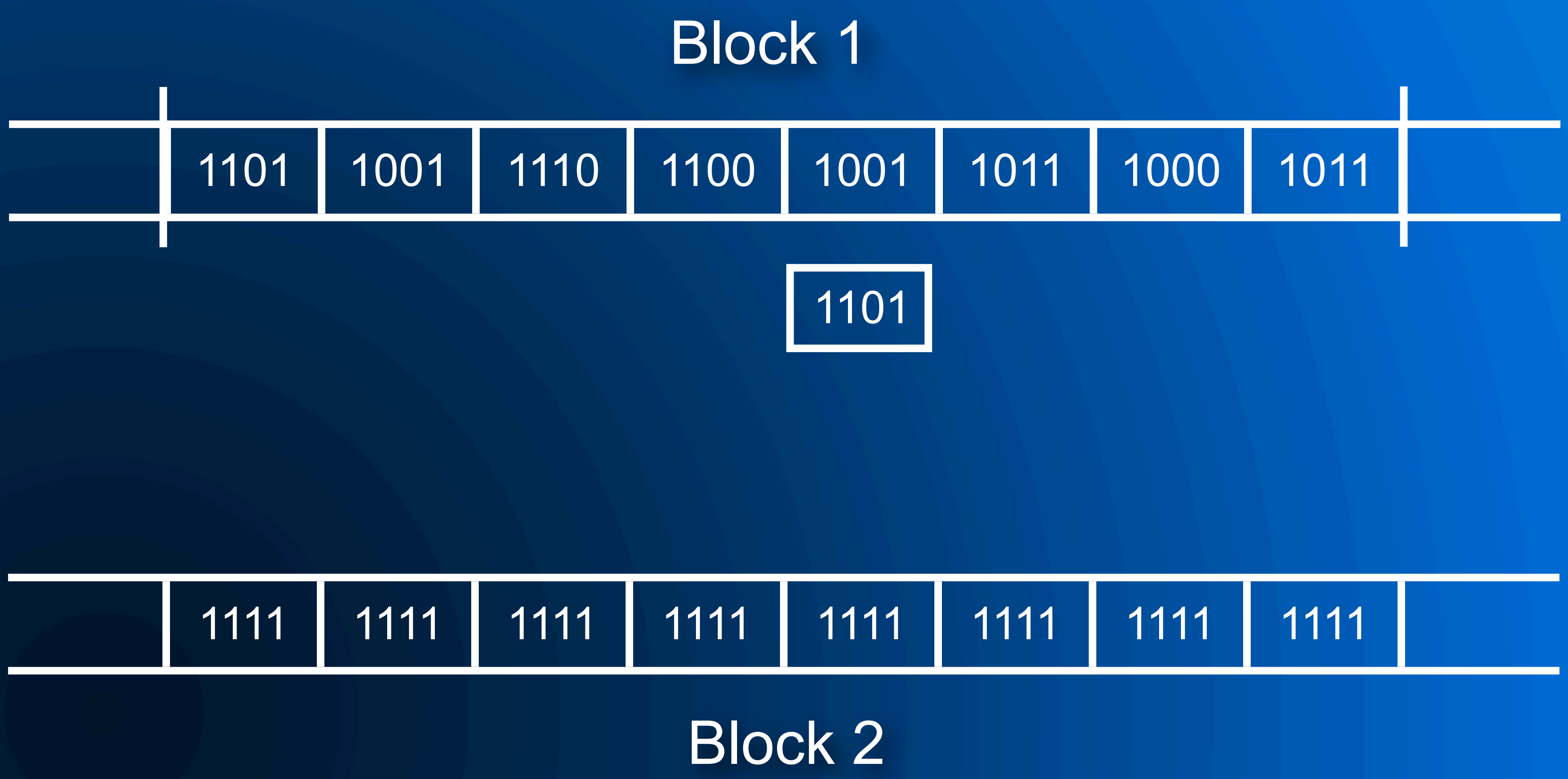
Architecture: Flash versus Disk

Problem: Page Updates



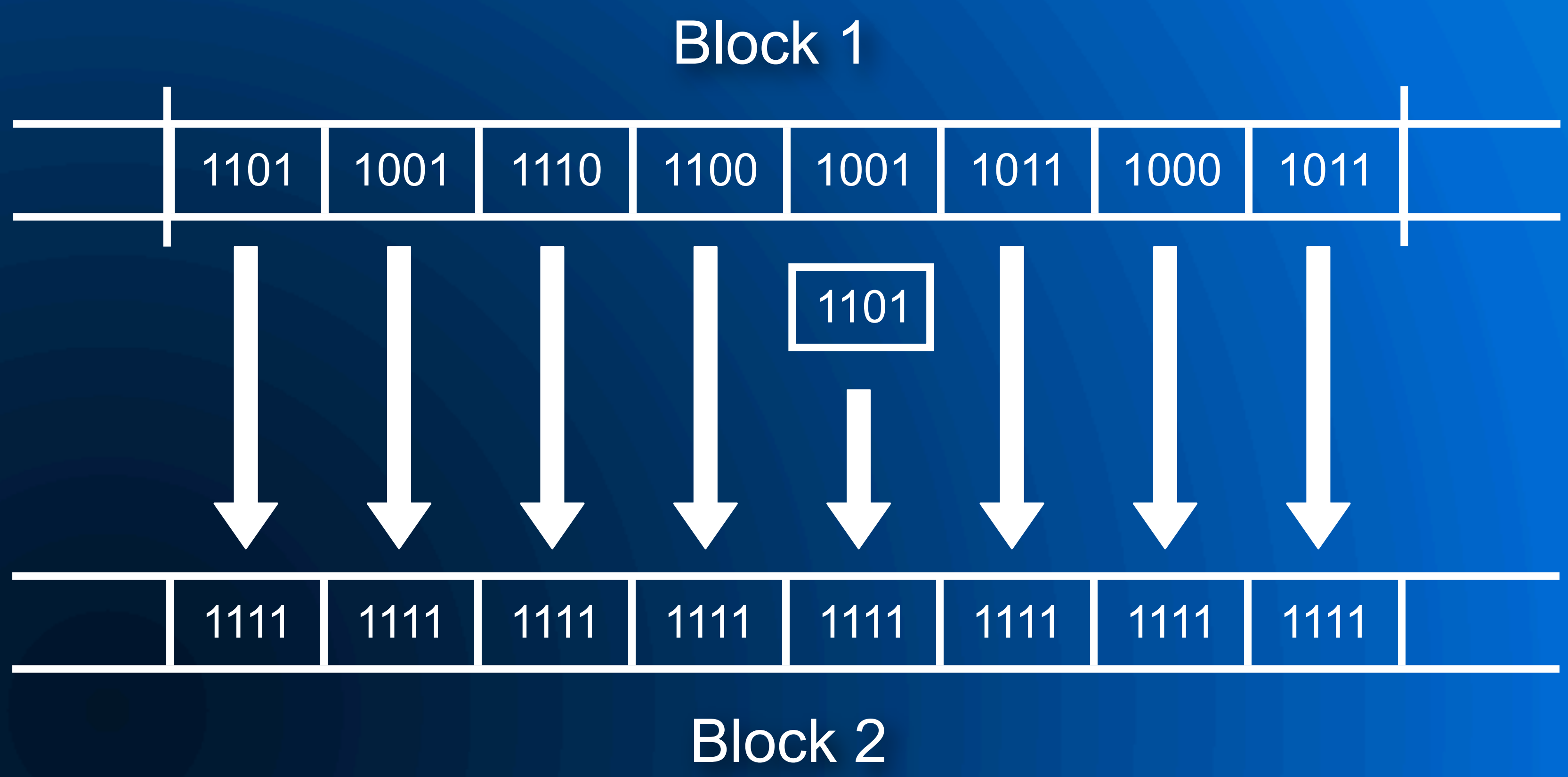
Architecture: Flash versus Disk

Problem: Page Updates



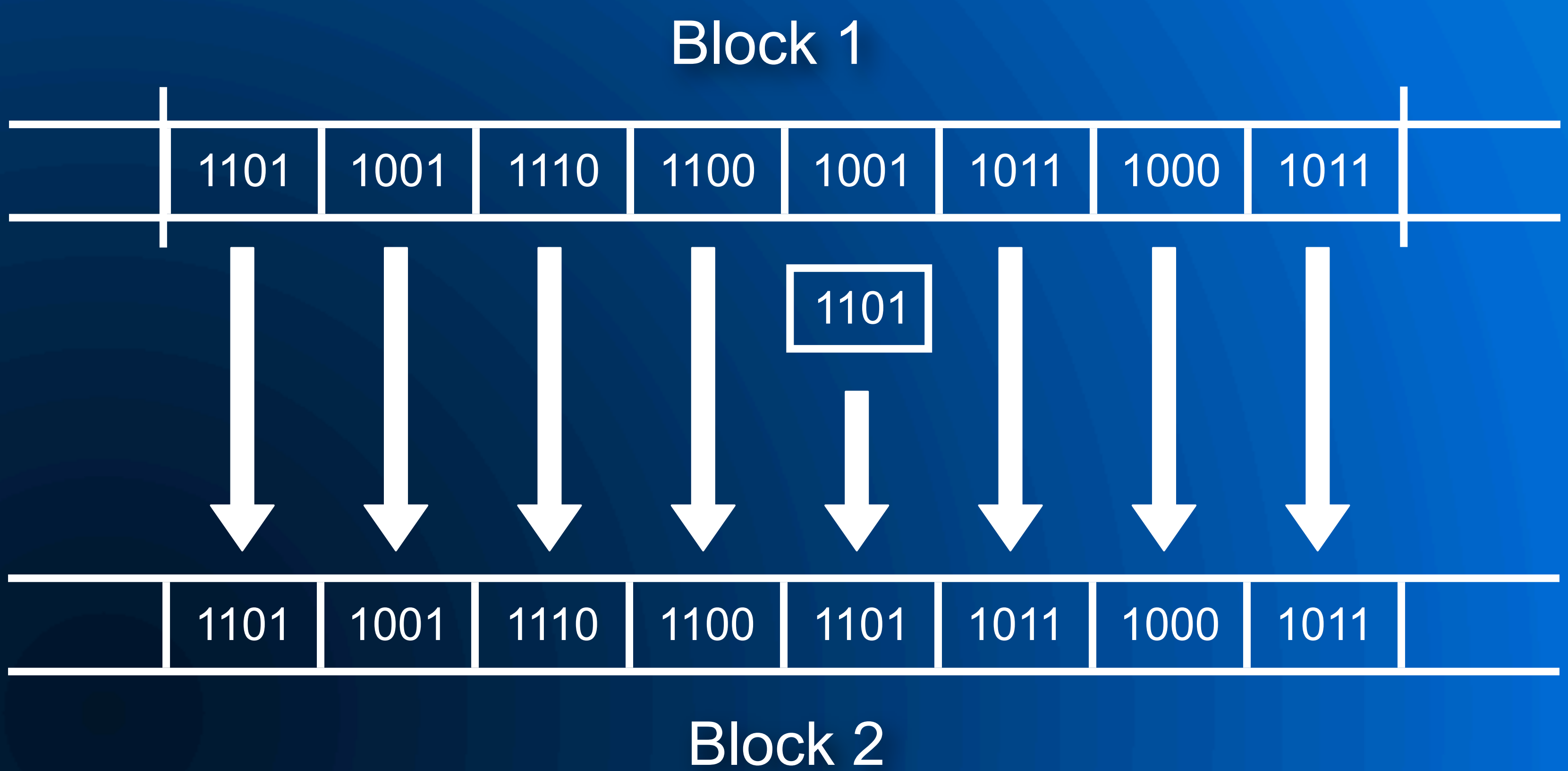
Architecture: Flash versus Disk

Problem: Page Updates



Architecture: Flash versus Disk

Problem: Page Updates



Architecture: Flash versus Disk

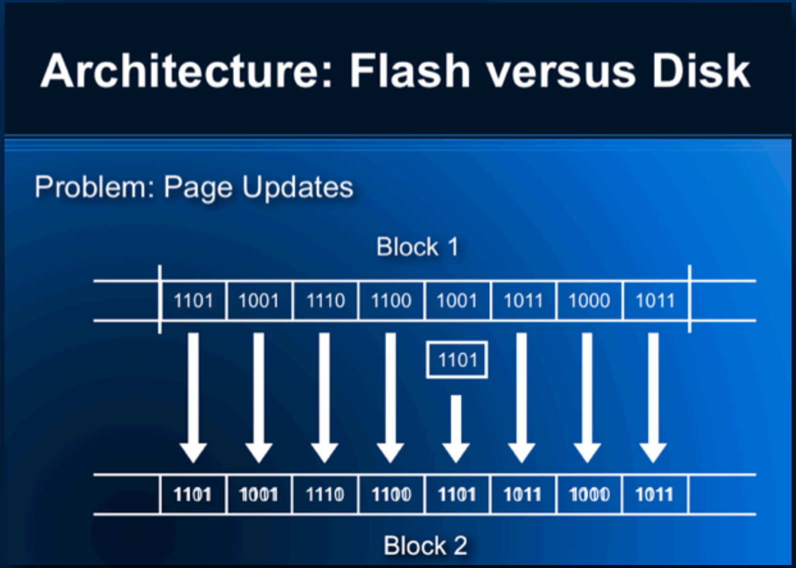
Problem: Write Endurance

Number of allowed erase cycles > 1.000.000 today

Architecture: Flash versus Disk

Problem: Write Endurance

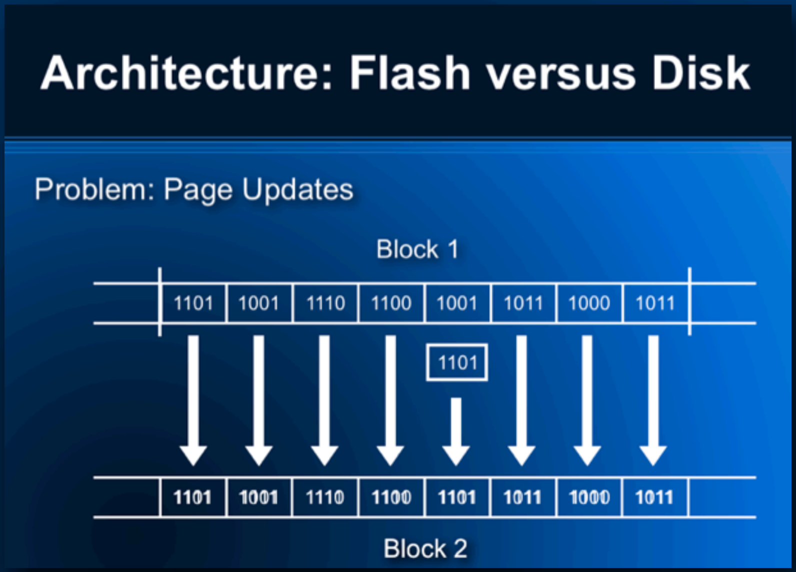
Number of allowed erase cycles > 1.000.000 today



Architecture: Flash versus Disk

Problem: Write Endurance

Number of allowed erase cycles > 1.000.000 today

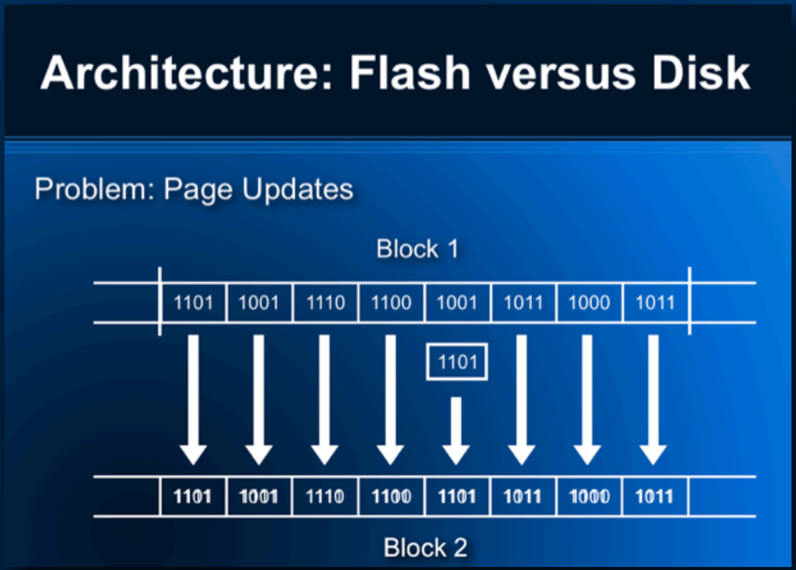


Uniform Block Distribution needed!

Architecture: Flash versus Disk

Problem: Write Endurance

Number of allowed erase cycles > 1.000.000 today



Uniform Block Distribution needed!

↓

Flash Translation Layer (FTL)

Architecture: Flash versus Disk

Flash Translation Layer

Wear Leveling

Guarded Secrets!

Architecture: Flash versus Disk

Flash Translation Layer

Wear Leveling

Guarded Secrets!

Problem: Hard to predict SSD Performance

Setup: Access Pattern

Sequential Access:

Random Access:

Random Sorted Access:

Setup: Access Pattern

Sequential Access:

Random Access:

Random Sorted Access:

Setup: Access Pattern

Sequential Access:

Random Access:

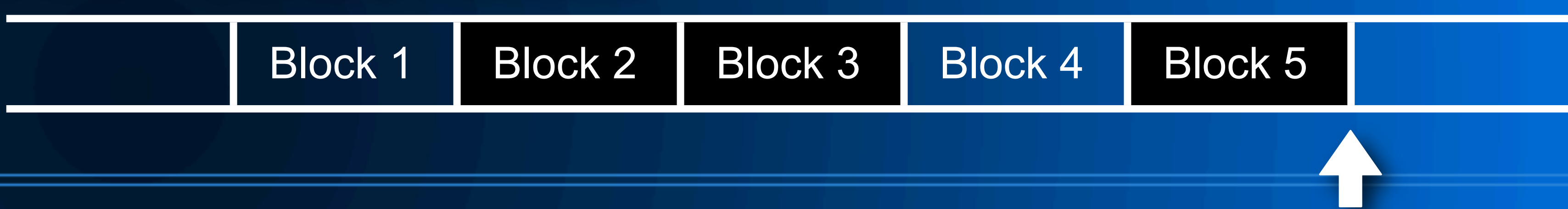
Random Sorted Access:

Setup: Access Pattern

Sequential Access:

Random Access:

Random Sorted Access:



Access Time: Flash versus Disk

HDD:

$$t_d(b) \approx t_{SIO} + t_{Savg} + t_{rev} / 2 + b / F_d$$

t_{Savg}	Average Seek Time
t_{rev}	Rotational Delay
F	Transfer Rate
t_{SIO}	CPU Overhead

Access Time: Flash versus Disk

HDD:

$$t_d(b) \approx t_{SIO} + t_{savg} + t_{rev} / 2 + b / F_d$$

SSD:

$$t_f(b) \approx t_{SIO} + b / F_f$$

t_{savg}	Average Seek Time
t_{rev}	Rotational Delay
F	Transfer Rate
t_{SIO}	CPU Overhead

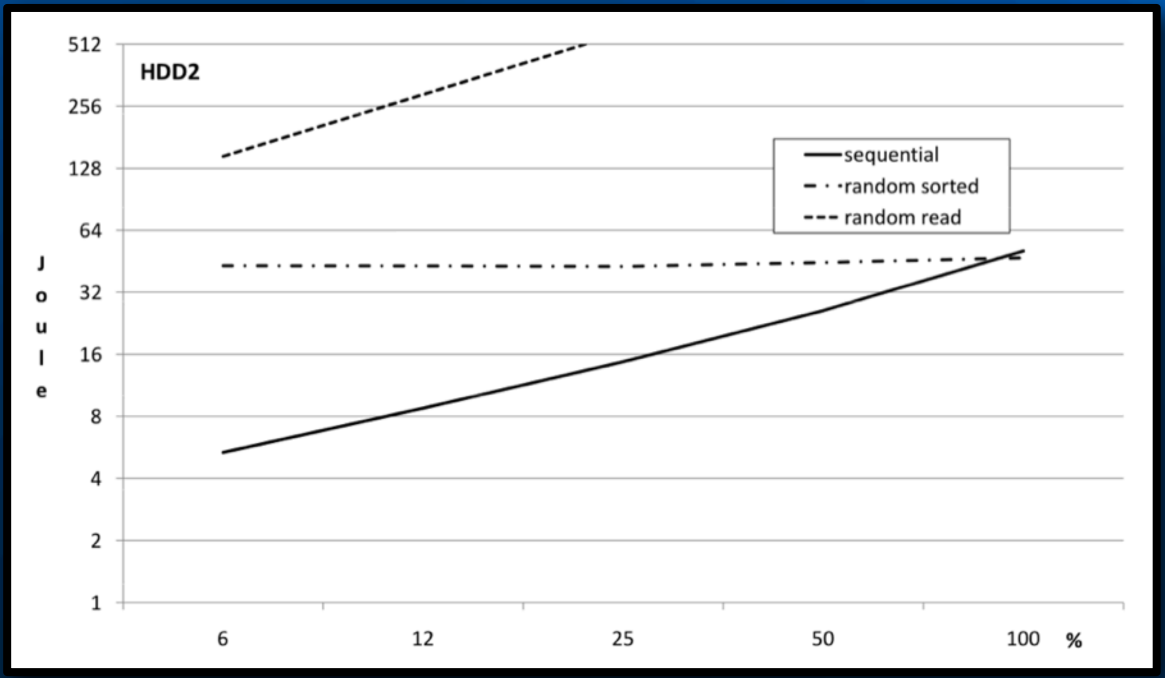
HDDs under Test

Device Type	Name	Idle Power Watt	Price $ / GB	Capacity	Rotation Speed
Maxtor IDE	HDD1	7.1	0.7	80 GB	7200 rpm
WD SATA	HDD2	4.5	1	150 GB	10000 rpm
Fujitsu SAS	HDD3	12.8	1.2	147 GB	15000 rpm

SSDs under Test

Device Type	Name	Idle Power Watt	Price $ / GB	Capacity	Rotation Speed
SuperTalent	SSD1	no info	4.3	32 GB	-
Mtron	SSD2	1.4	17.6	32 GB	-
Intel	SSD3	0.1	2.9	160 GB	-

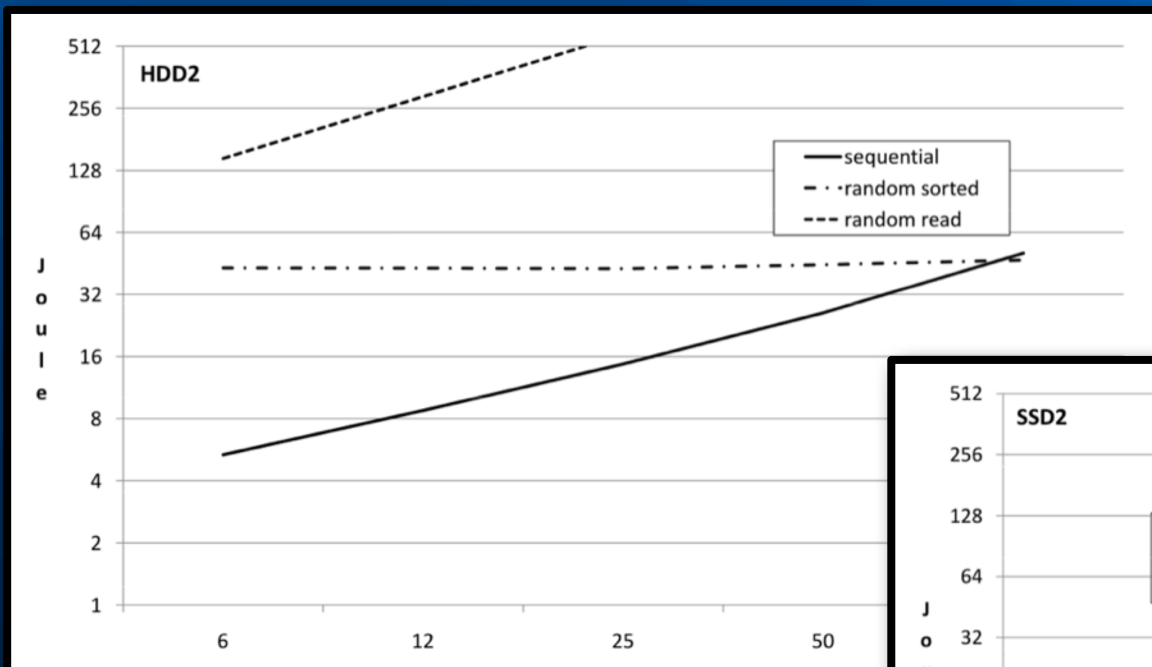
Energy Cost: Flash versus Disk



HDD

% of file read

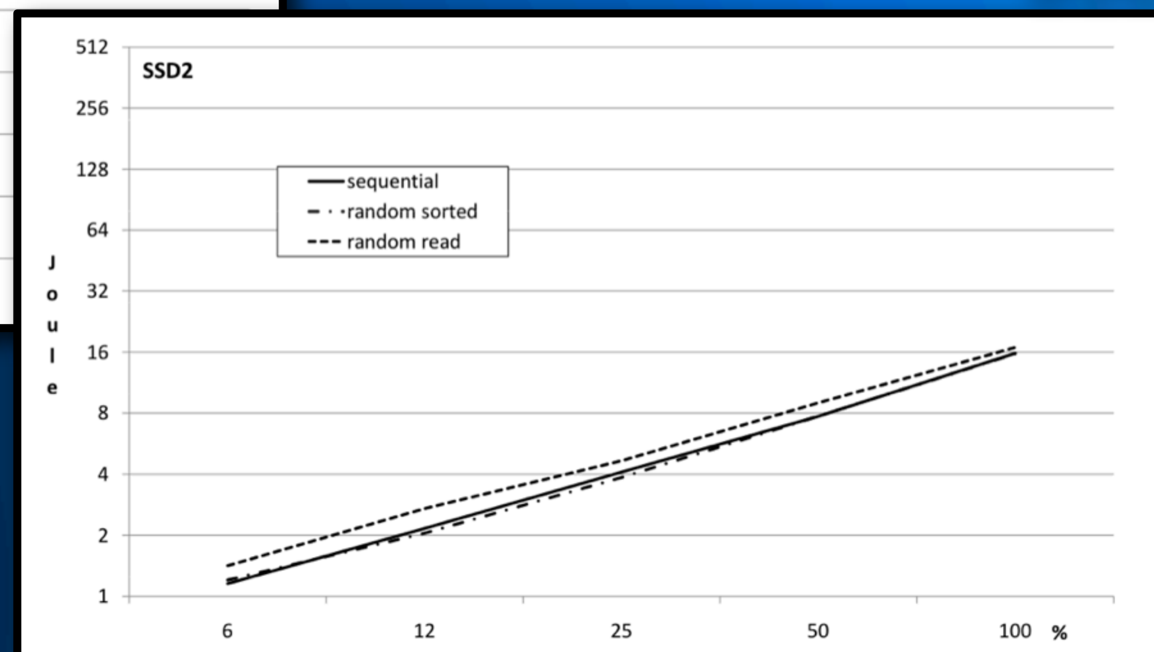
Energy Cost: Flash versus Disk



HDD

% of file read

SSD



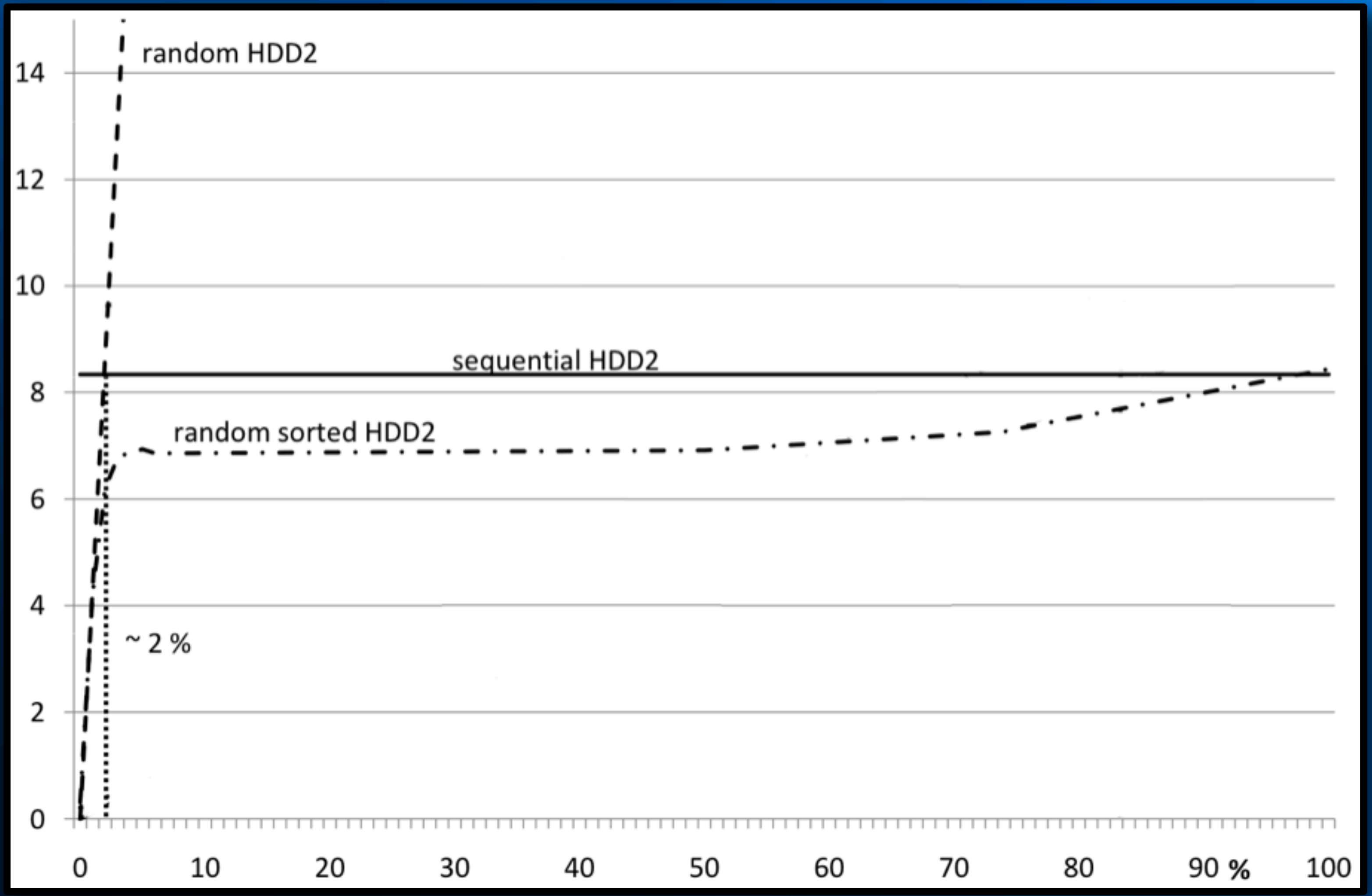
Random vs. Sequential Access

Rule of Thumb (for HDDs):

If > 2% of file accessed → Read complete file!

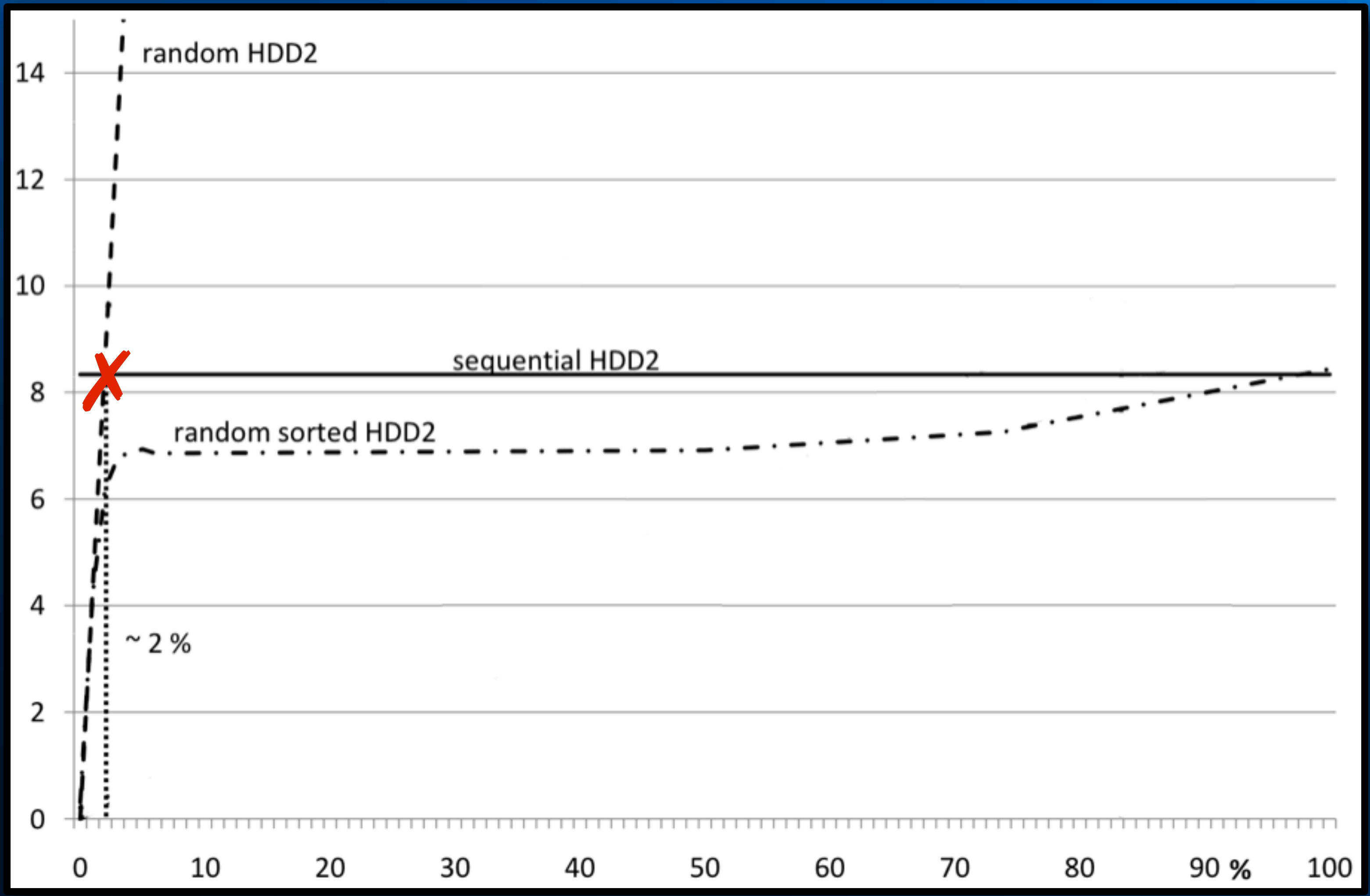
Random vs. Sequentiell Access

Time in sec



% of file read

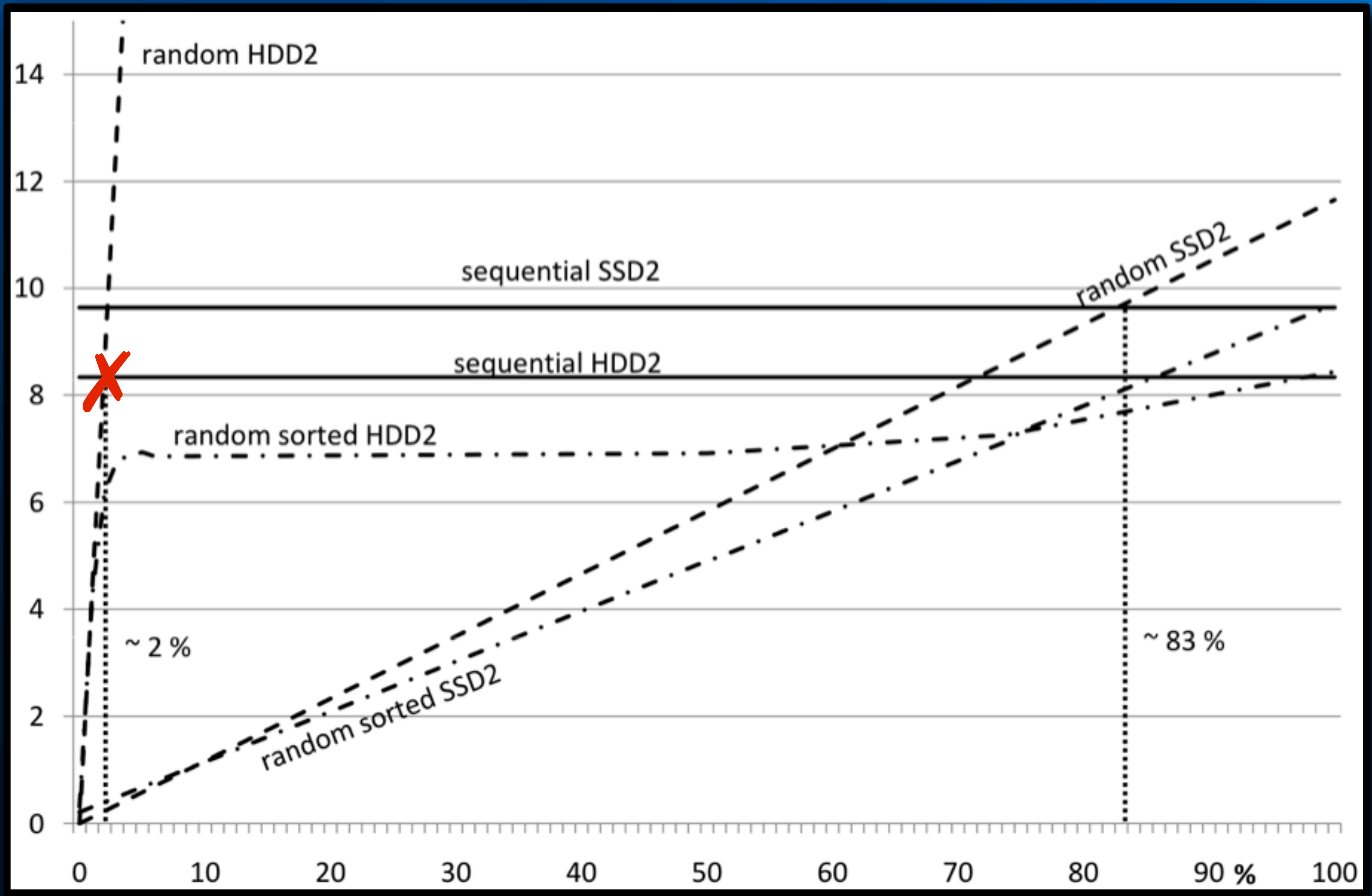
Random vs. Sequentiell Access



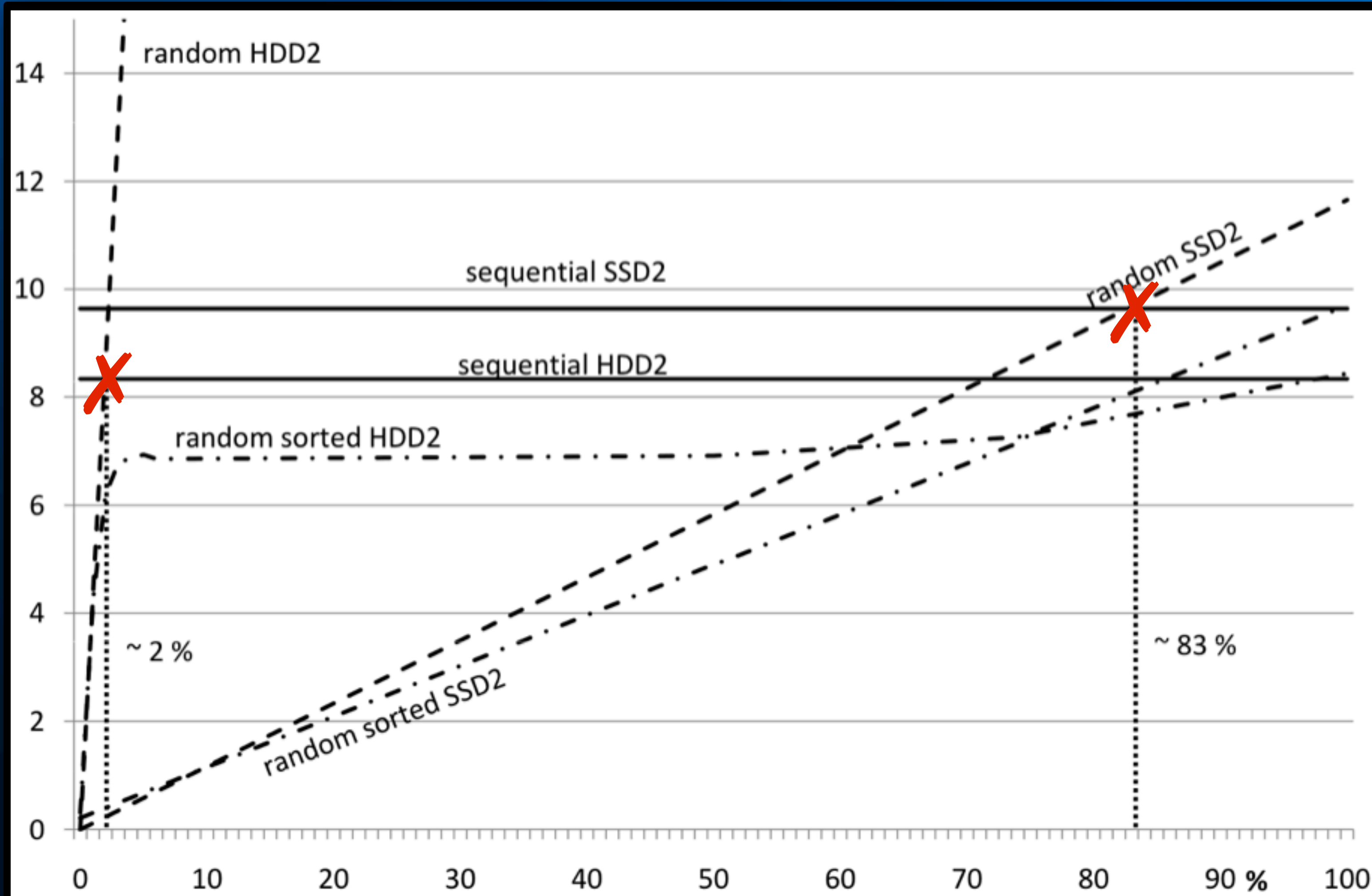
Time in sec

% of file read

Random vs. Sequentiell Access



Random vs. Sequentiell Access



Time in sec

% of file read

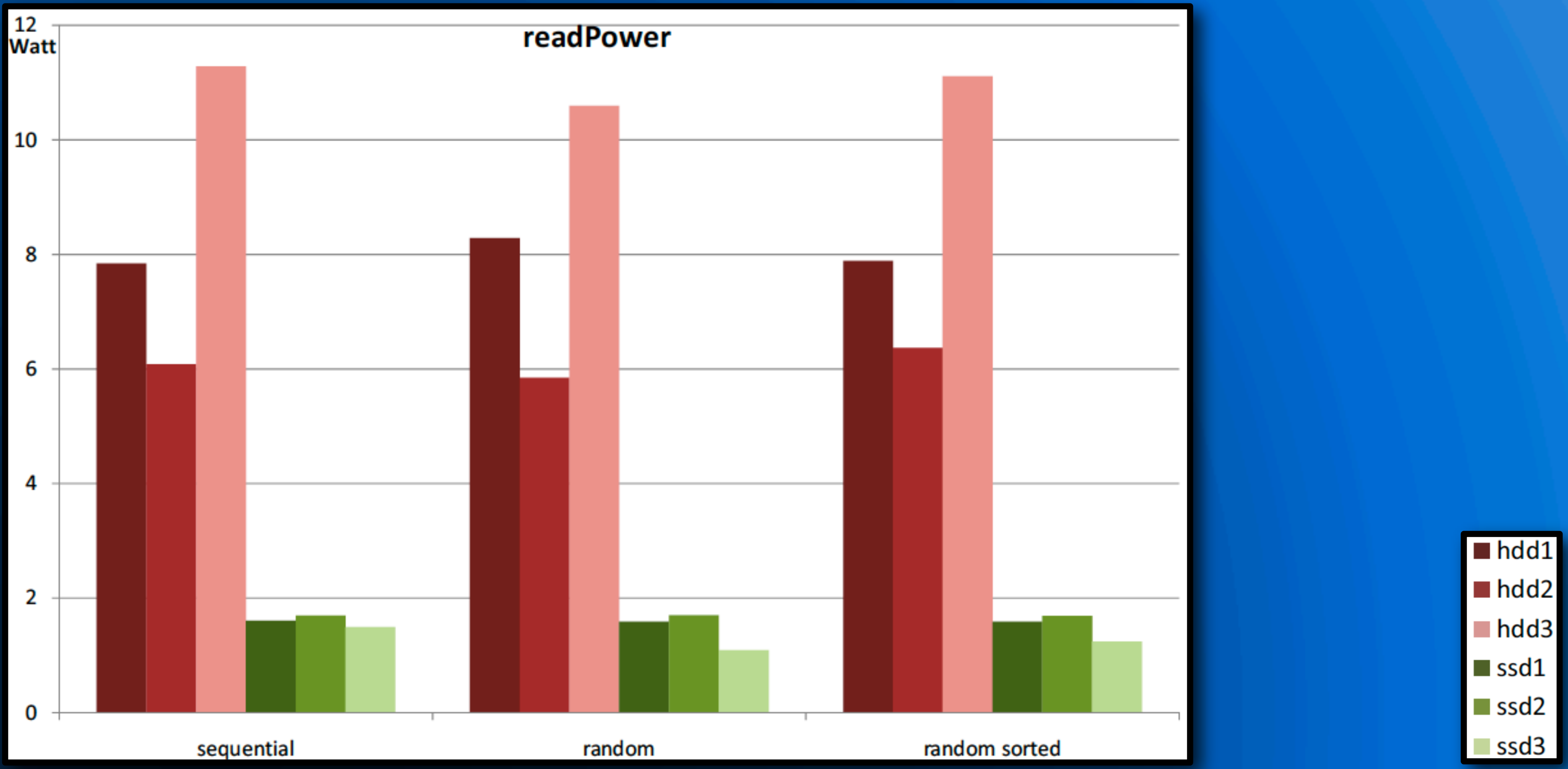
Random vs. Sequentiell Access

Rule of Thumb: Read complete file ...

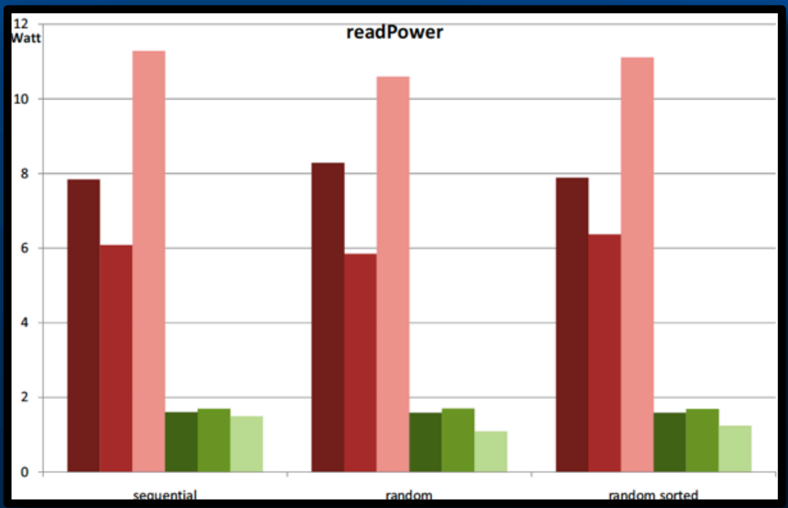
HDD: ... if > 2% of file accessed.

SSD: ... if > 83% of file accessed.

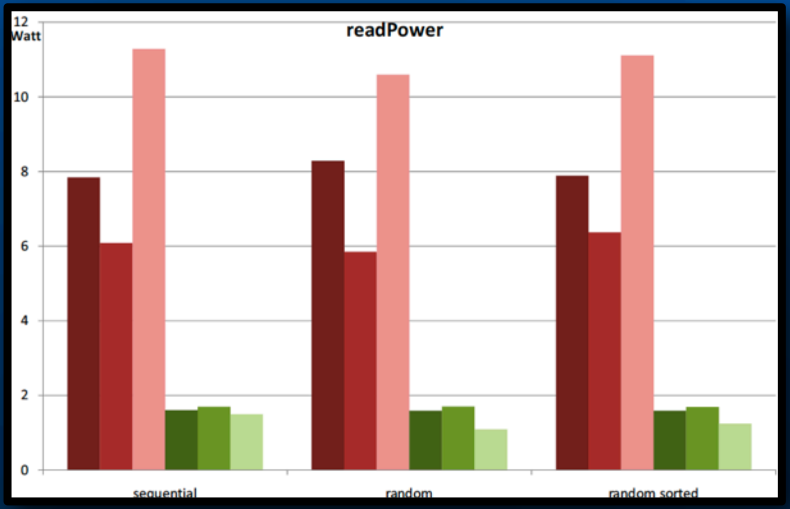
Experiment: Power Consumption



Experiment: Power Consumption

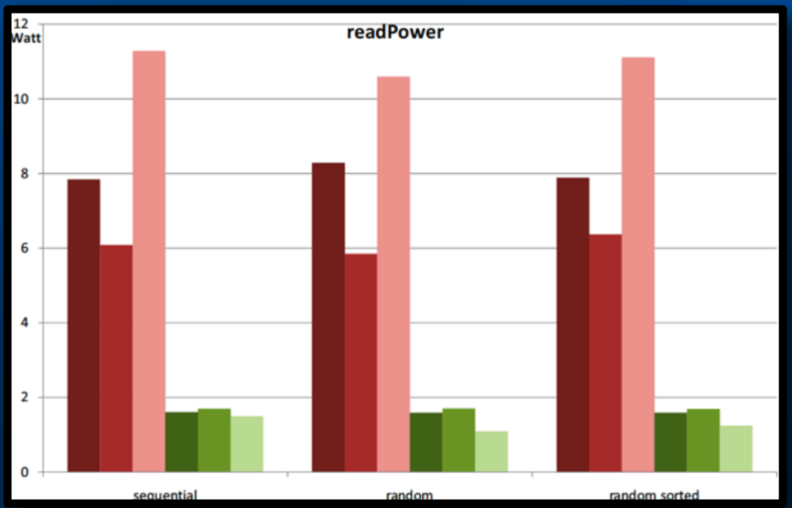


Experiment: Power Consumption



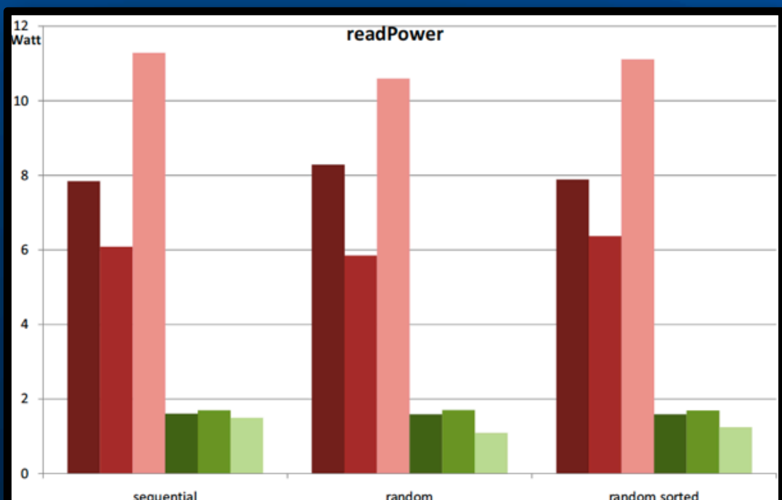
- HDDs consume more power in all cases.

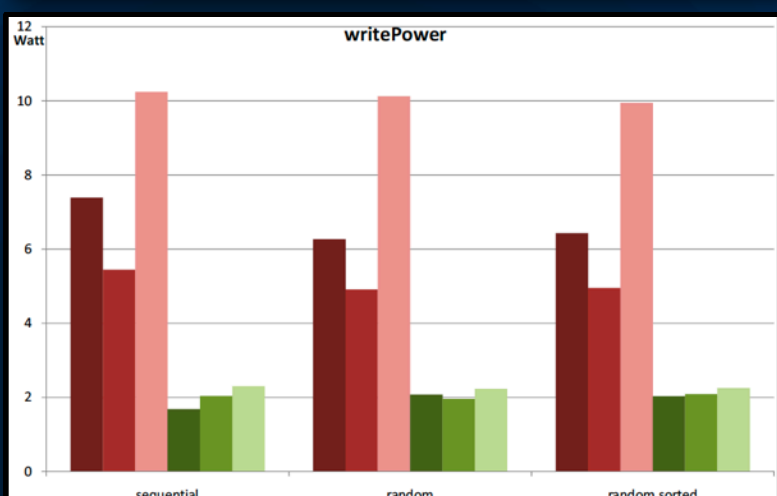
Experiment: Power Consumption



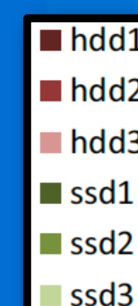
- HDDs consume more power in all cases.
- Constant consumption for SSDs

Experiment: Power Consumption

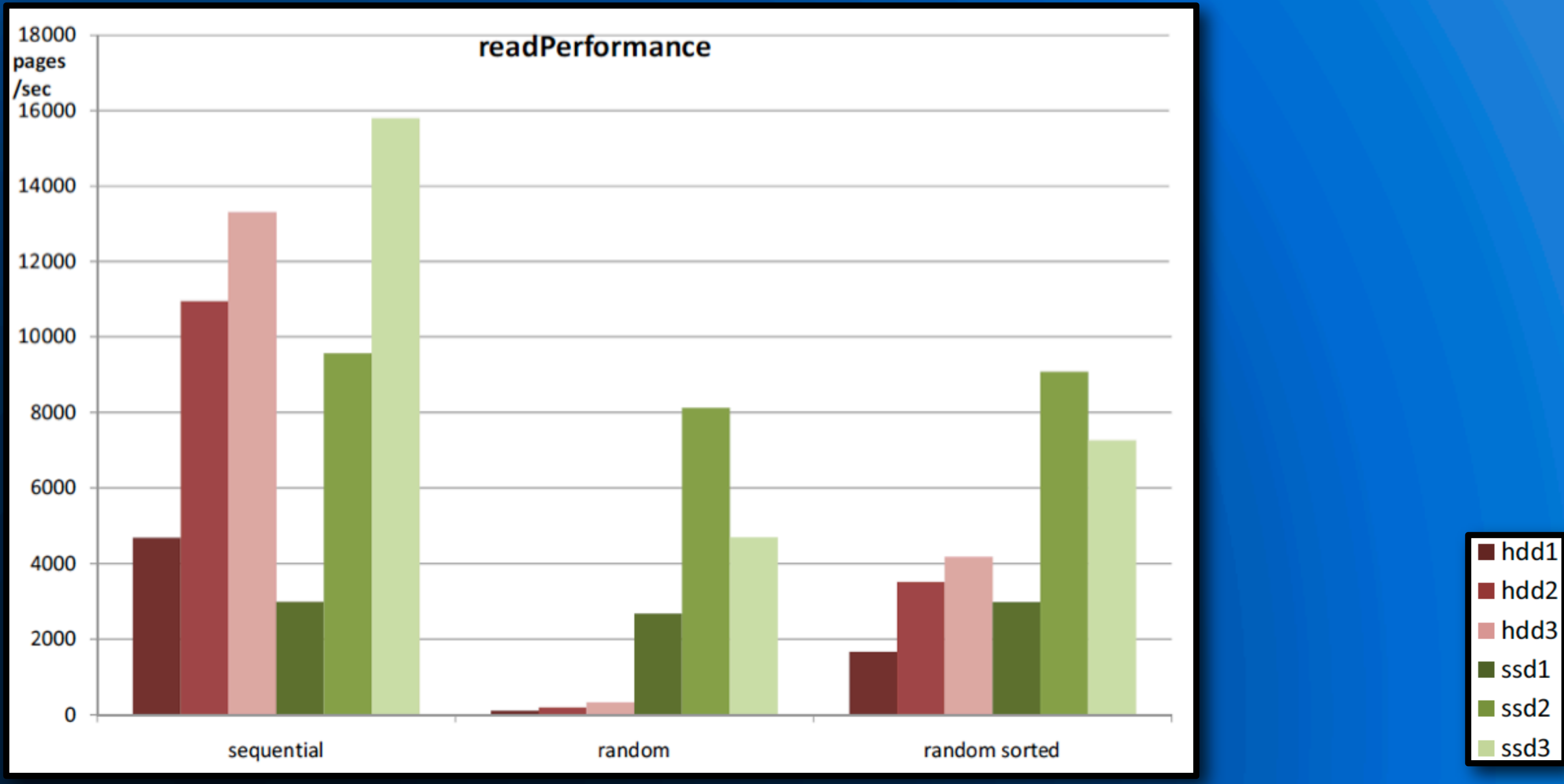




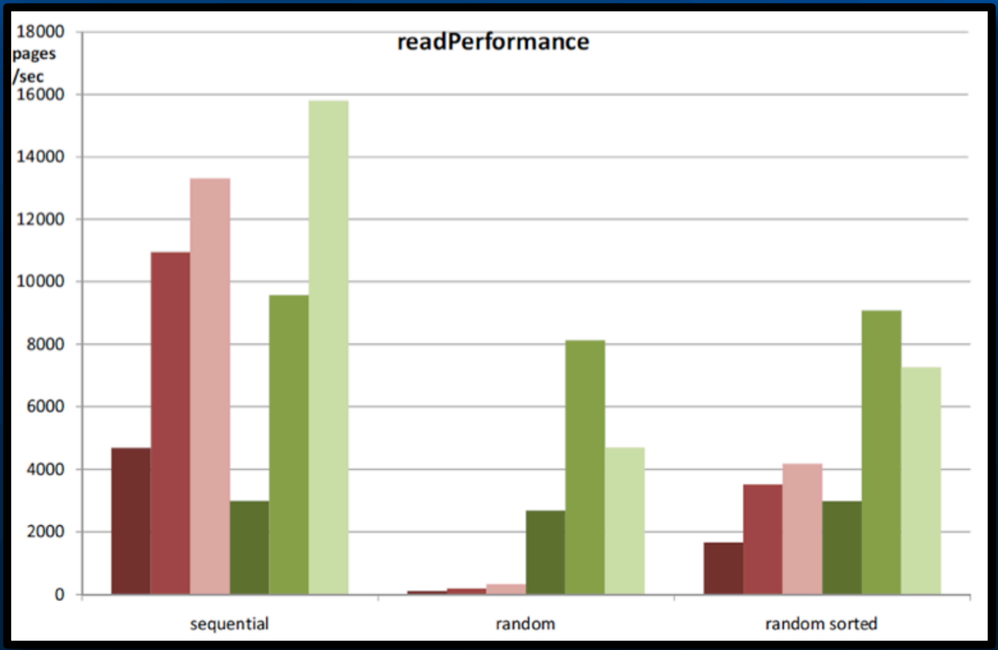
- HDDs consume more power in all cases.
- Constant consumption for SSDs
- Read/Write has marginal impact



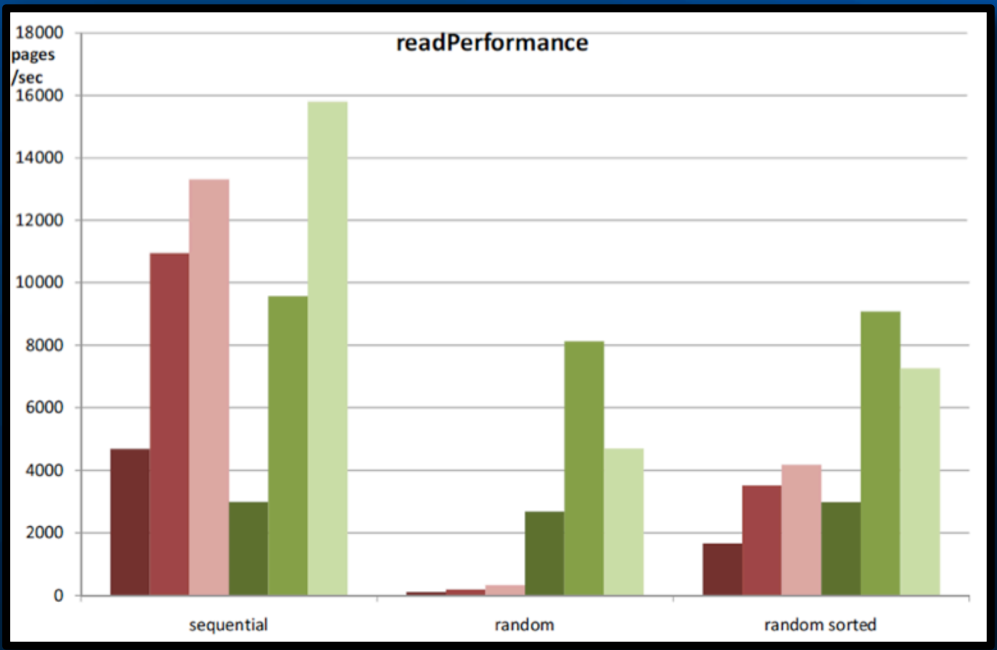
Experiment: Performance



Experiment: Performance

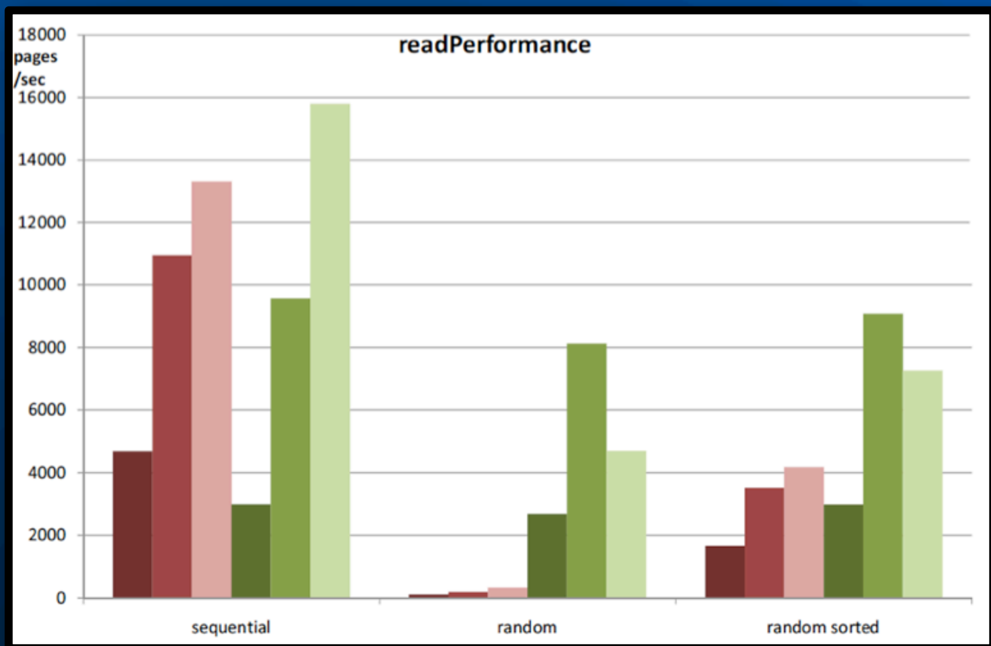


Experiment: Performance



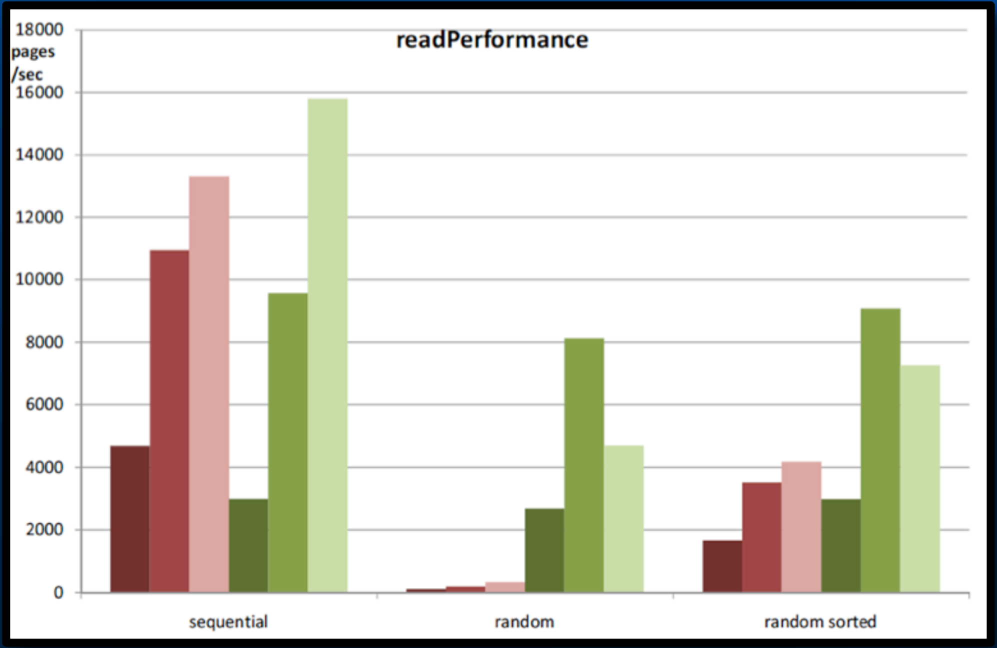
- Sequential performance comparable.

Experiment: Performance



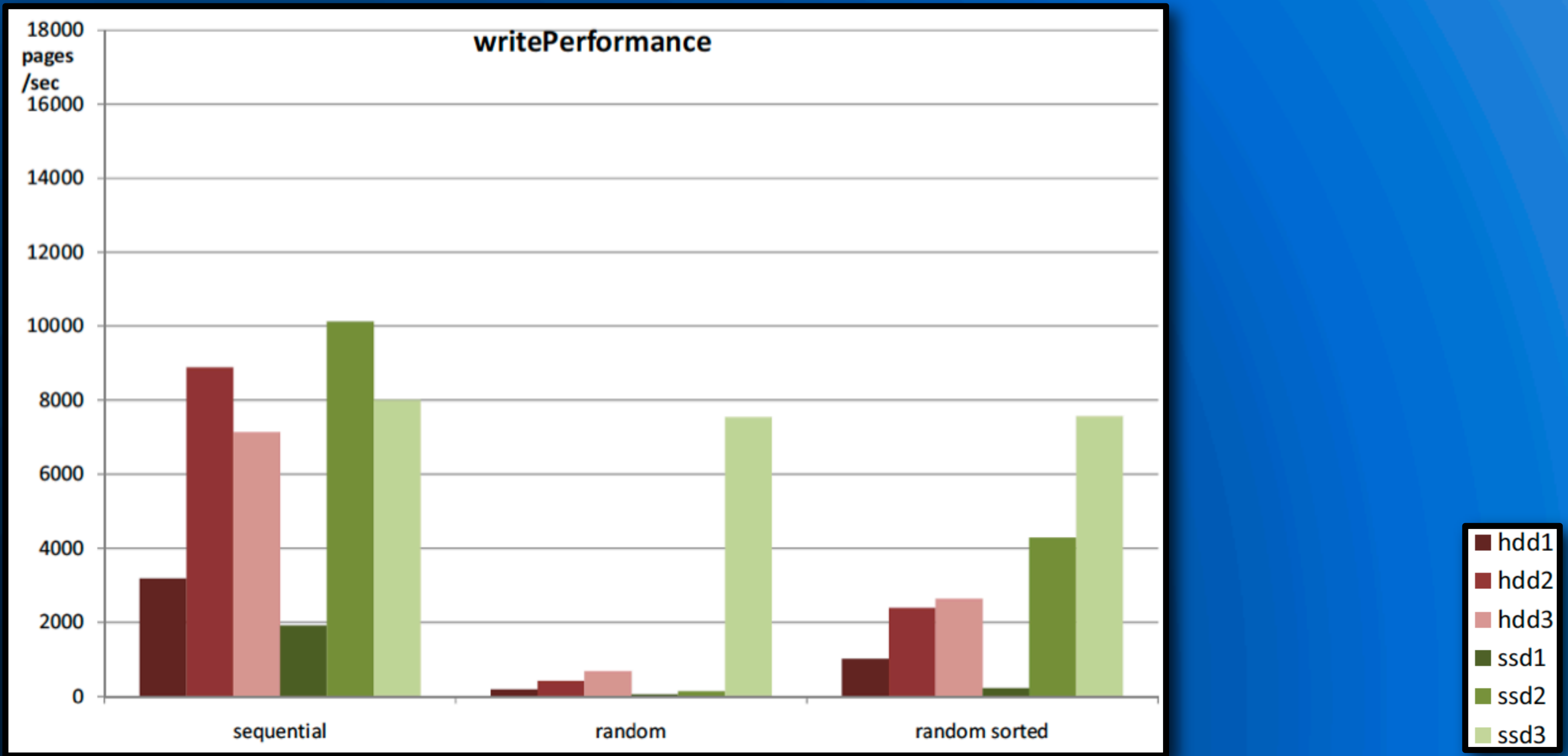
- Sequential performance comparable.
- Random read is worst case for HDDs.

Experiment: Performance

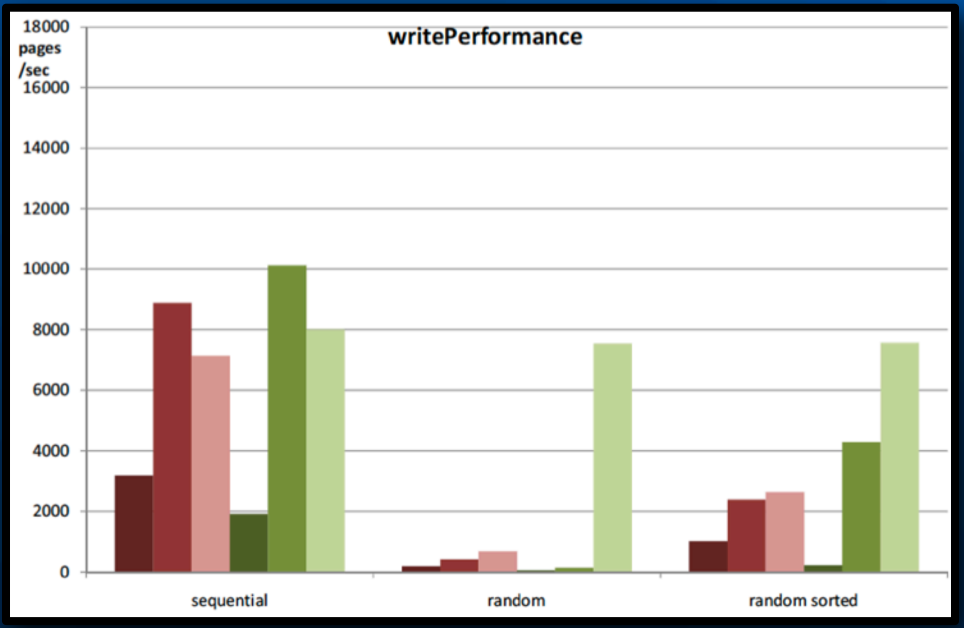


- Sequential performance comparable.
- Random read is worst case for HDDs.
- Sorting decreases this problem.

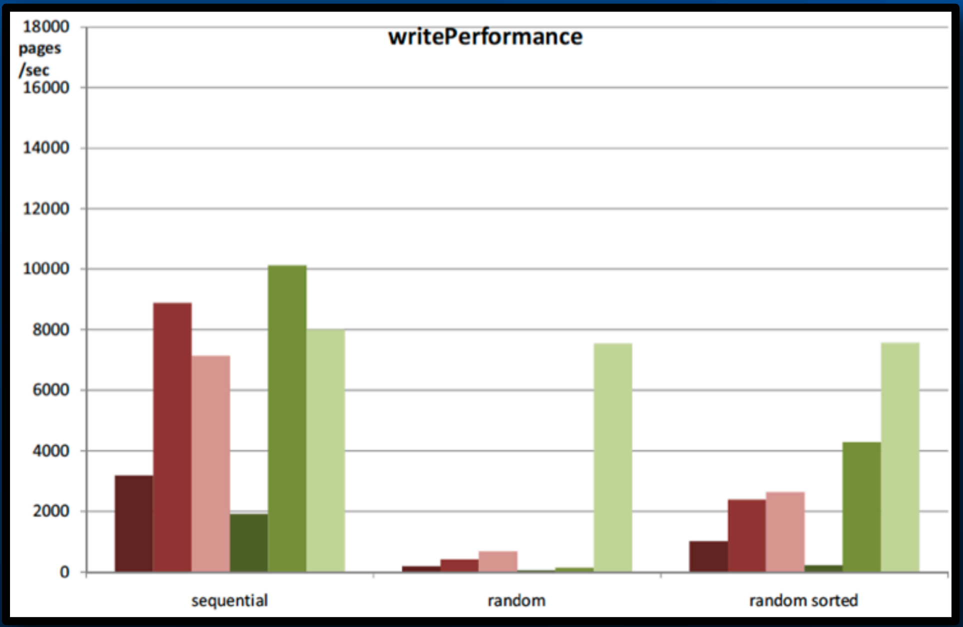
Experiment: Performance



Experiment: Performance

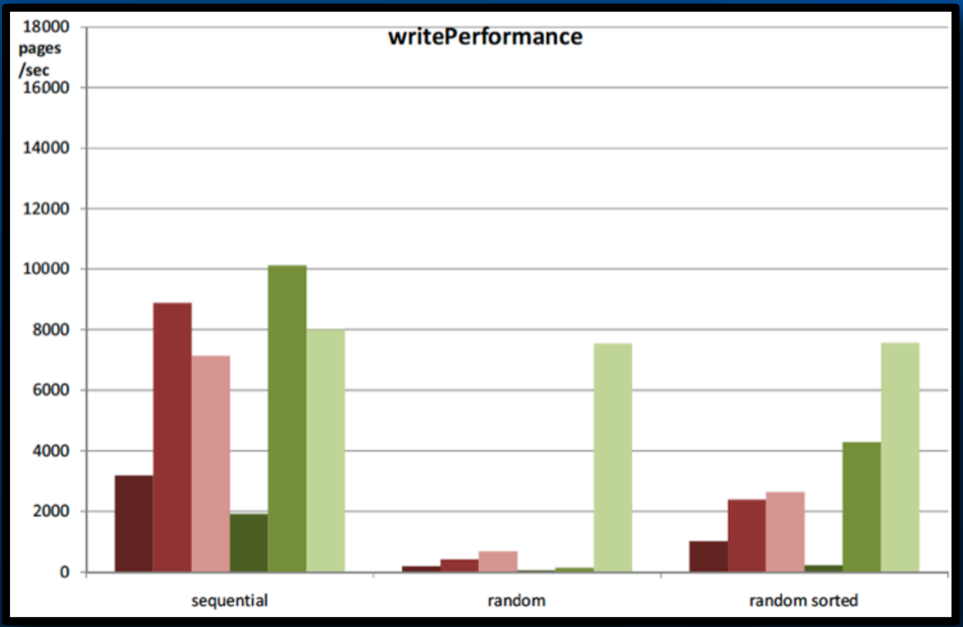


Experiment: Performance



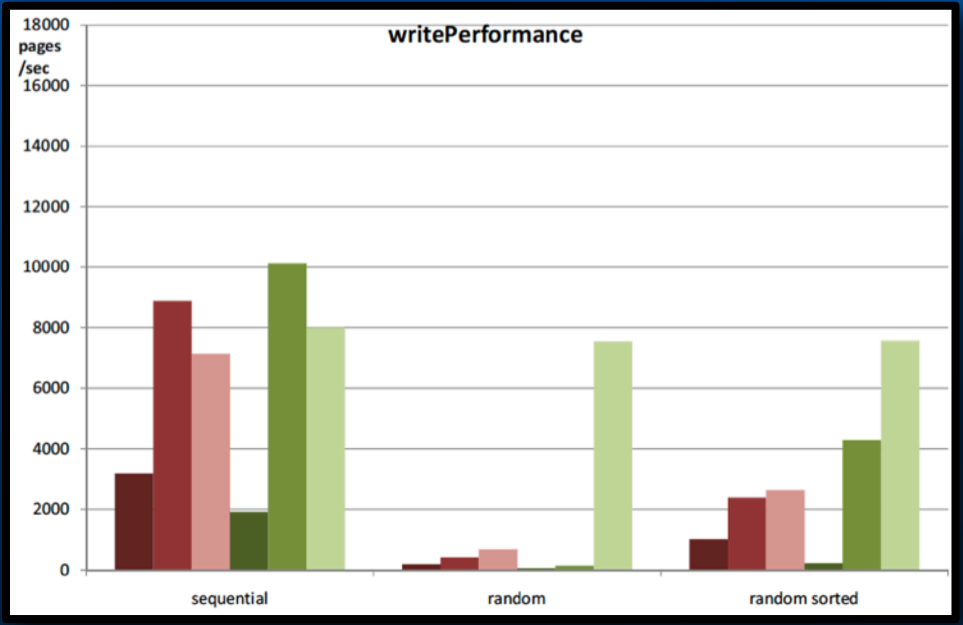
- Sequential performance comparable.

Experiment: Performance



- Sequential performance comparable.
- Random write problem for all, except of SSD3.

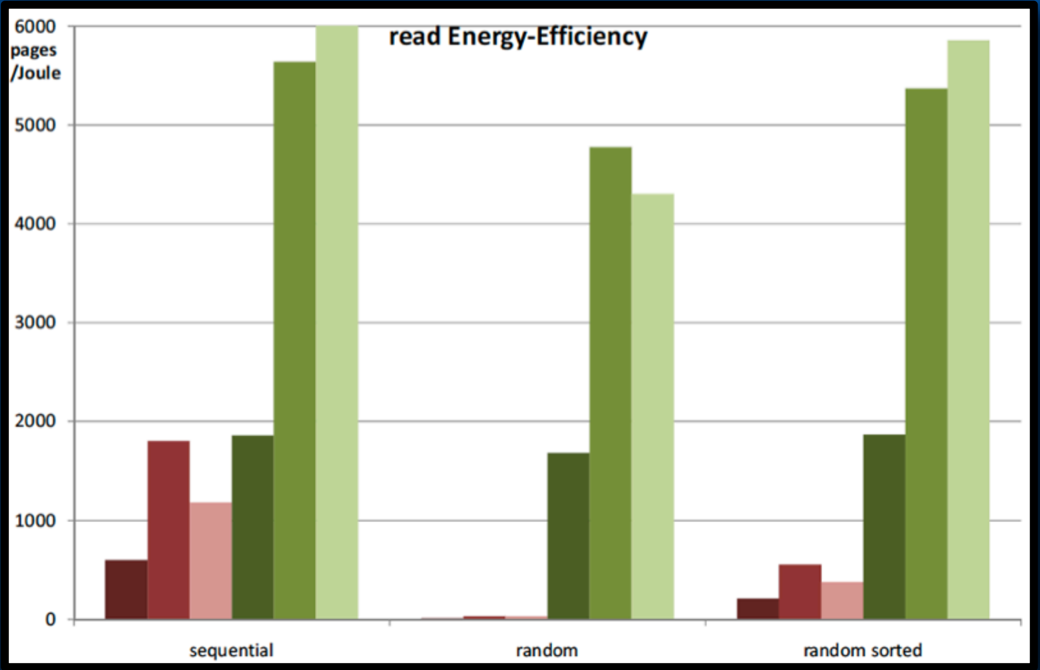
Experiment: Performance



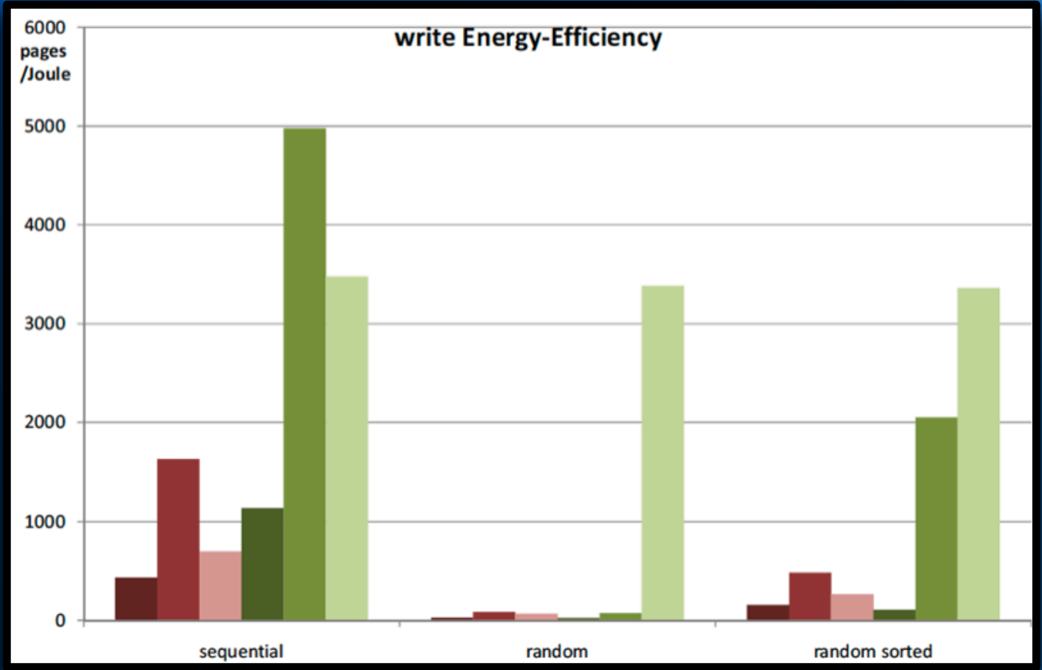
- Sequential performance comparable.
- Random write problem for all, except of SSD3.
- Sorting decreases this problem.

Experiment: Energy Efficiency

Read



Write



SSDs superior in ALL situations.

SSDs superior in MOST situations.

Drive Diversity

Main factors for performance and power consumption:

Drive Diversity

Main factors for performance and power consumption:

Rotation Speed

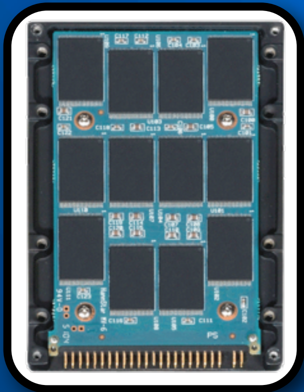
Capacity

Drive Diversity

Main factors for performance and power consumption:

Rotation Speed

Capacity



FTL

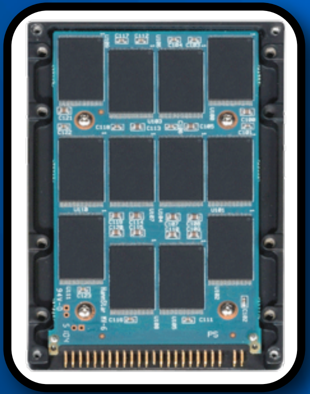
Unknown, proprietary firmware

Drive Diversity

Main factors for performance and power consumption:

Rotation Speed

Capacity



FTL

Unknown, proprietary firmware

Hardware vs. Software Specification

Flash inside of a DBMS

Till now: Direct I/O Mode ← pure file reading / writing

Flash inside of a DBMS

Till now: **Direct I/O Mode** ← pure file reading / writing

↓

Next step: **Integrate SSDs into DBMS**

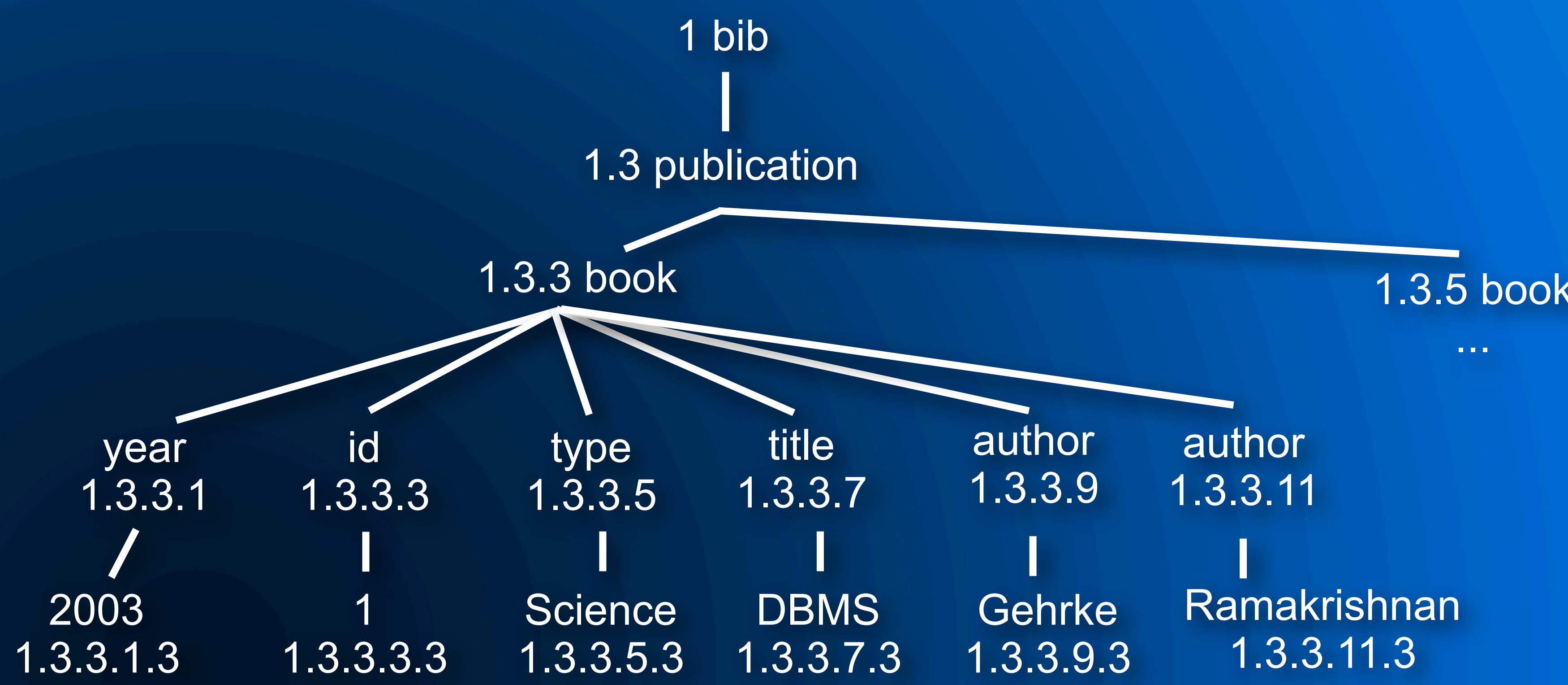
← index structures
← physical layout
← query types
← buffers / caches

XTC DBMS

XML DBMS: Stores tree-structured data

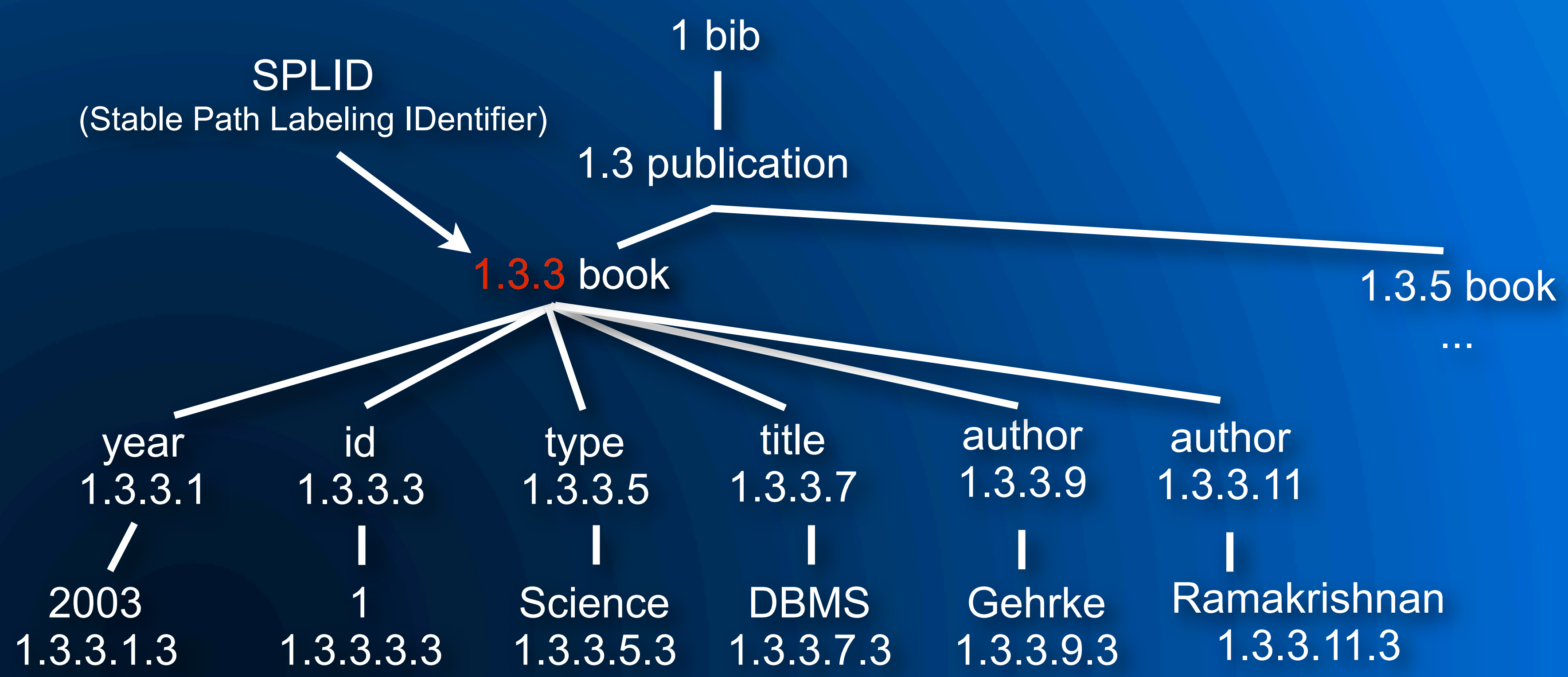
XTC DBMS

XML DBMS: Stores tree-structured data



XTC DBMS

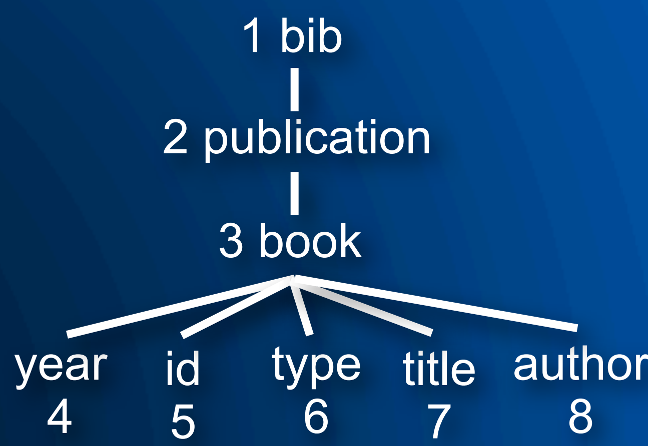
XML DBMS: Stores tree-structured data



XTC DBMS

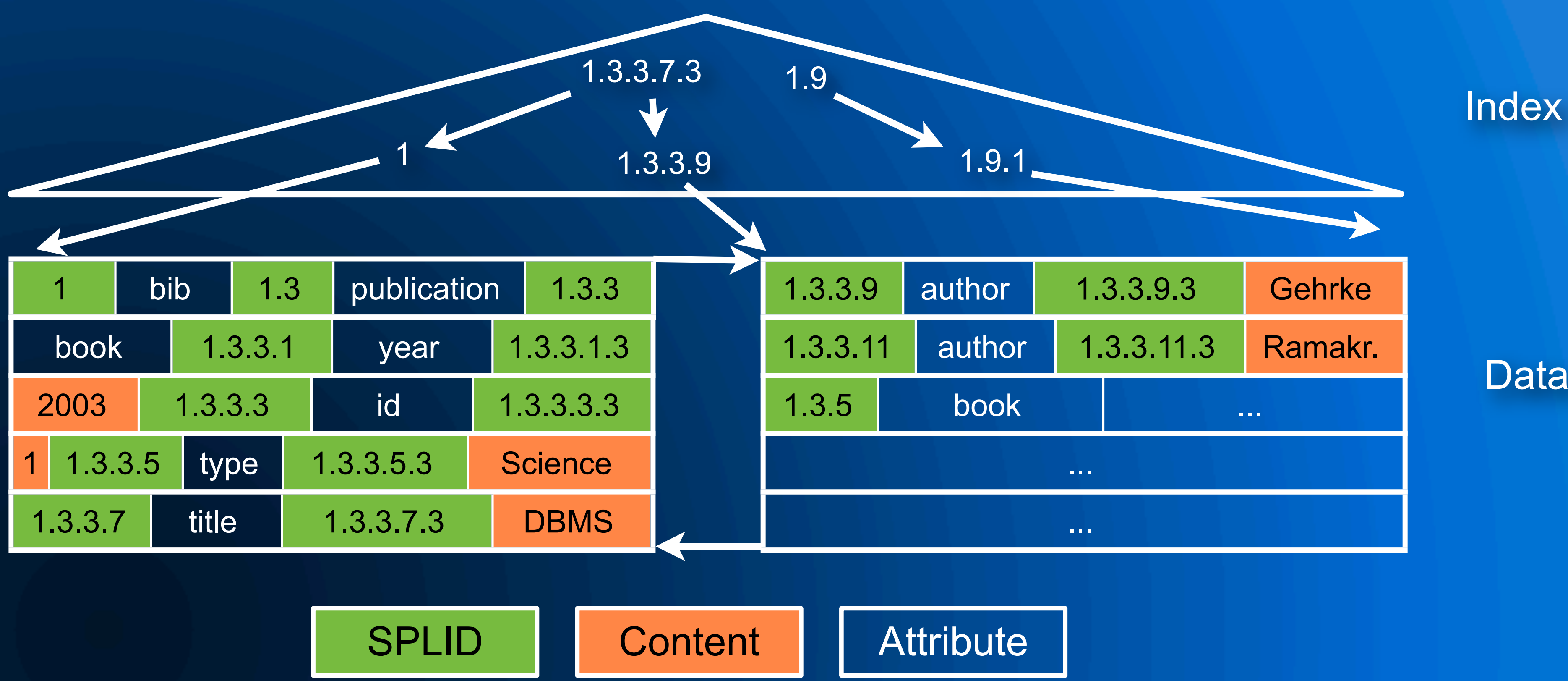
XML DBMS: Stores tree-structured data

Path Synopsis PCRs
(Path Class References)



XTC DBMS

XML DBMS: Physical representation



XTC DBMS

XML DBMS: Storage Mappings

- Full Representation (full) → previous slide — 95%
- Prefix-Compressed Representation (pc) — 70%

1	bib	1.3	publication	1.3.3

→

1	bib	3	publication	3

- Elementless Representation (eless) — 65%

1	bib	3	publication	3

→

1	1	3	2	3

of plain XML file

XTC DBMS - Optimizations

Problem:

DBMS and OS apply several optimizations.

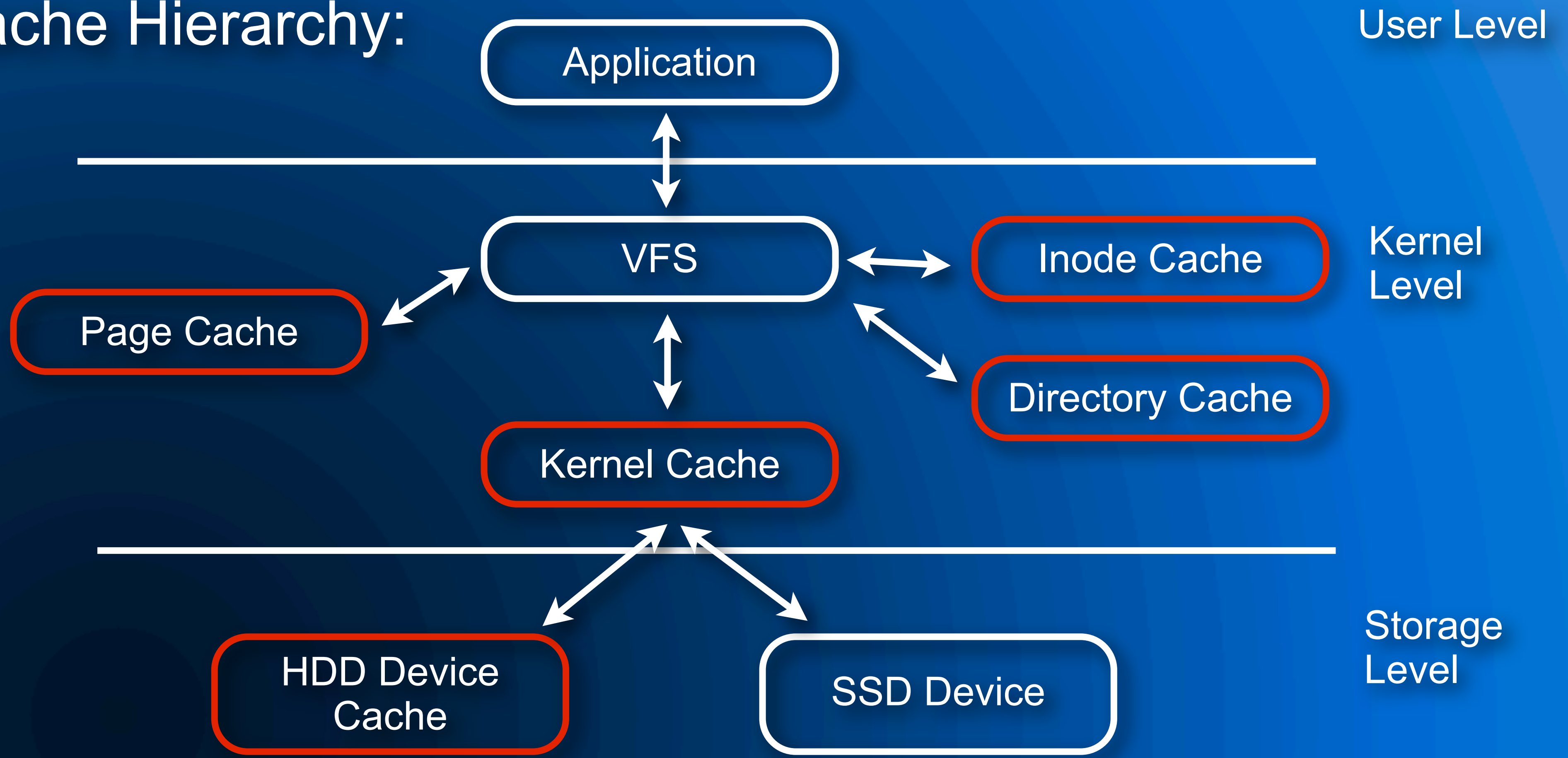
Good for Performance!

Bad for Experiments!

Caches, Buffers and Prefetching may disturb results.

XTC DBMS - Optimizations

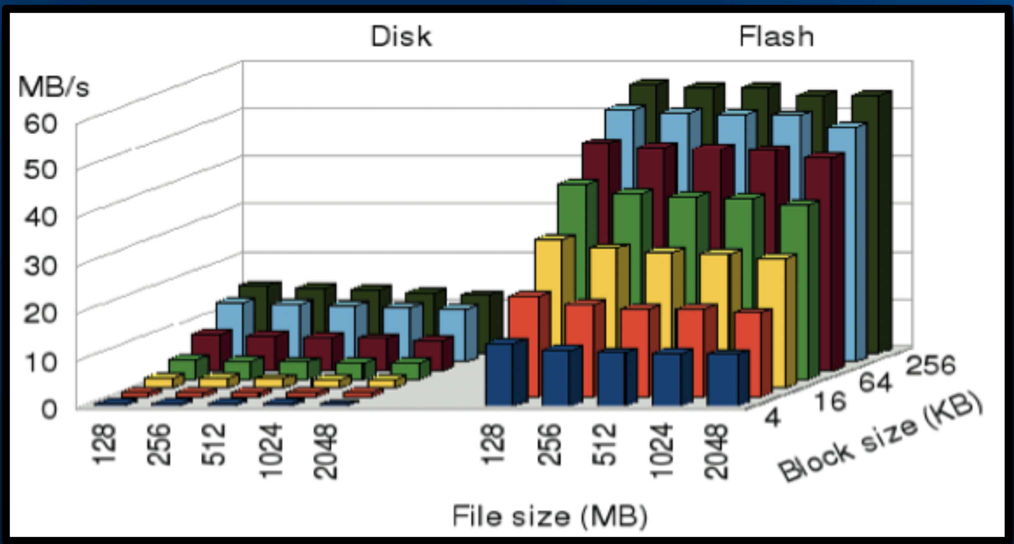
Cache Hierarchy:

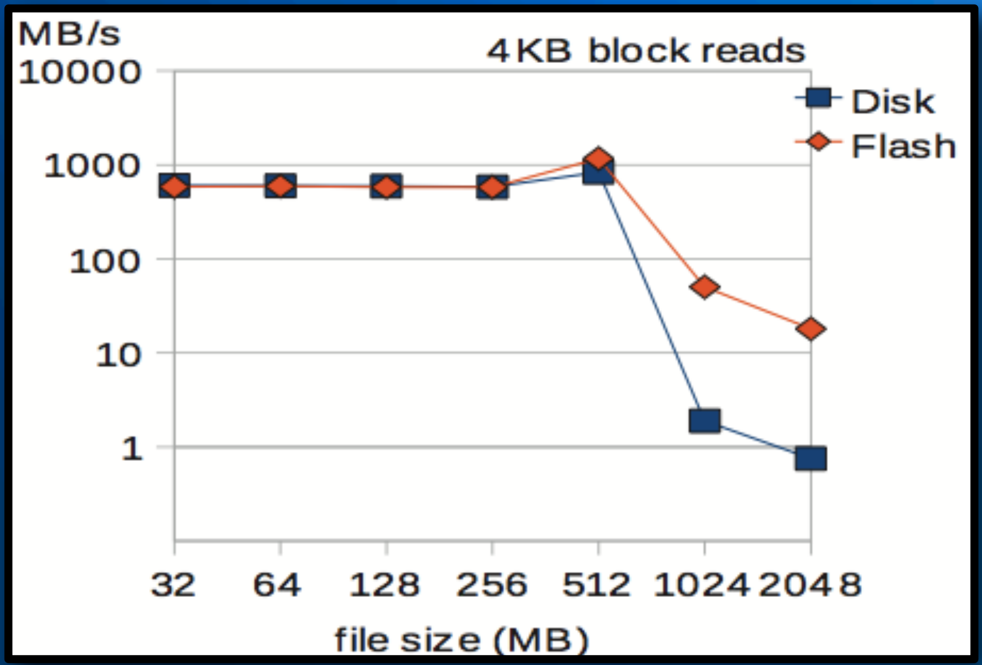


XTC DBMS - Optimizations

Problem:

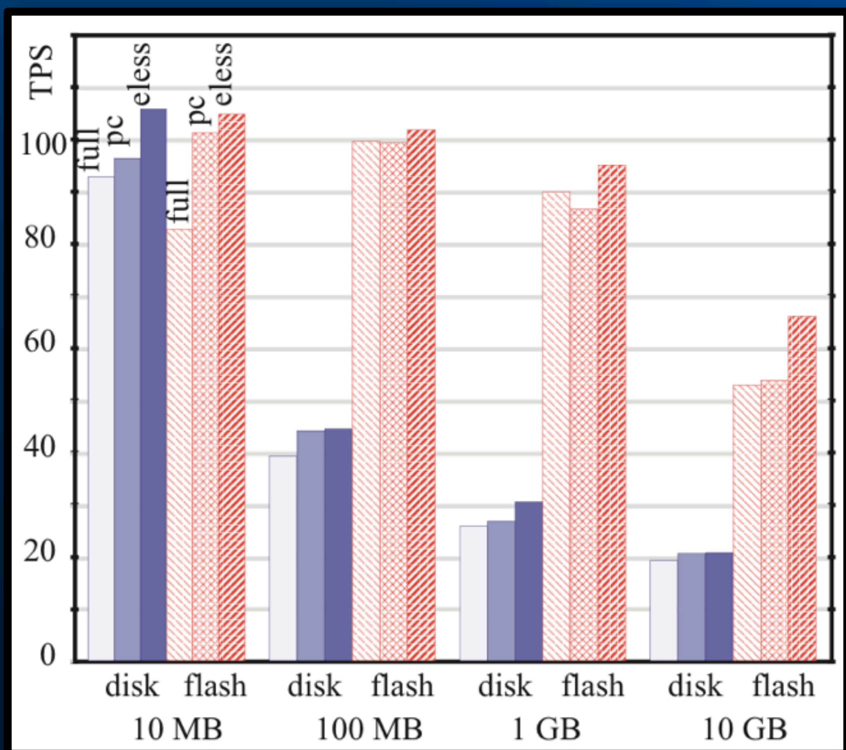
Cold Buffers vs. Hot Buffers



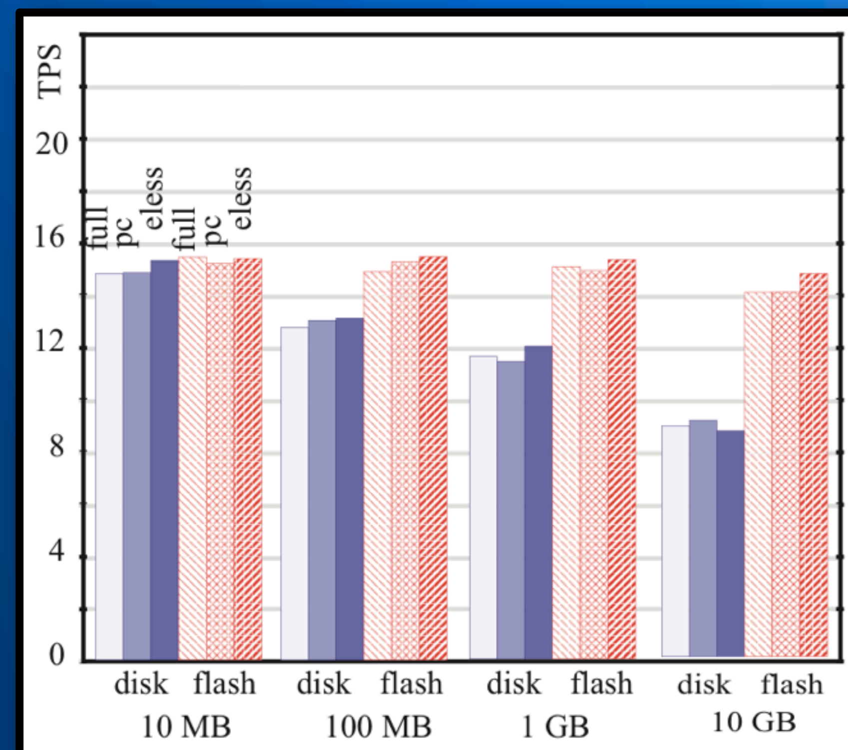


XTC DBMS - Experiments

Read performance: 1000 Transactions



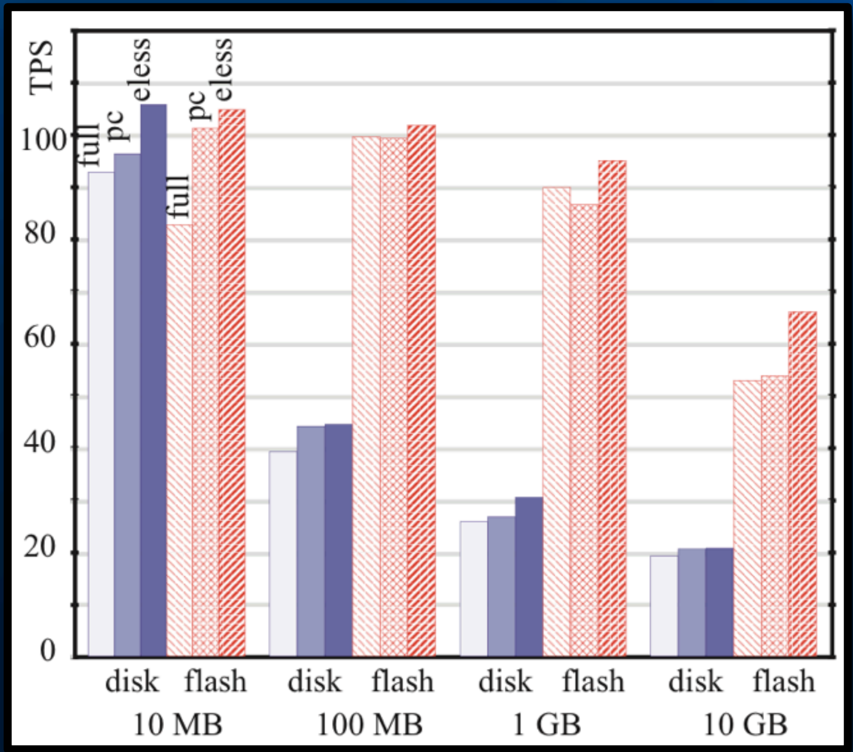
I/O bound

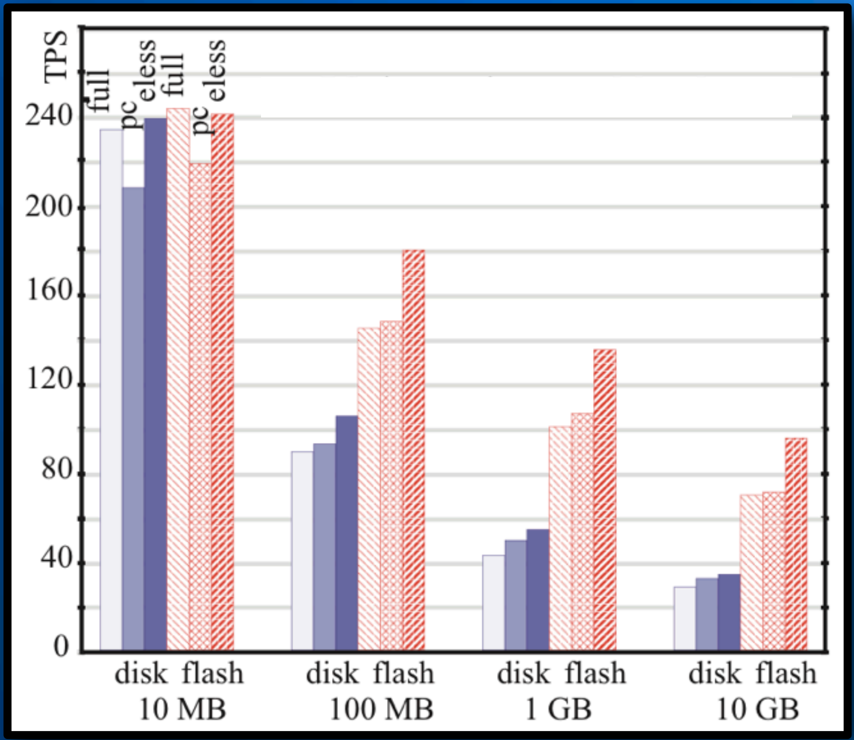


CPU bound

XTC DBMS - Experiments

Read performance: 1000 vs. 10000 Transactions

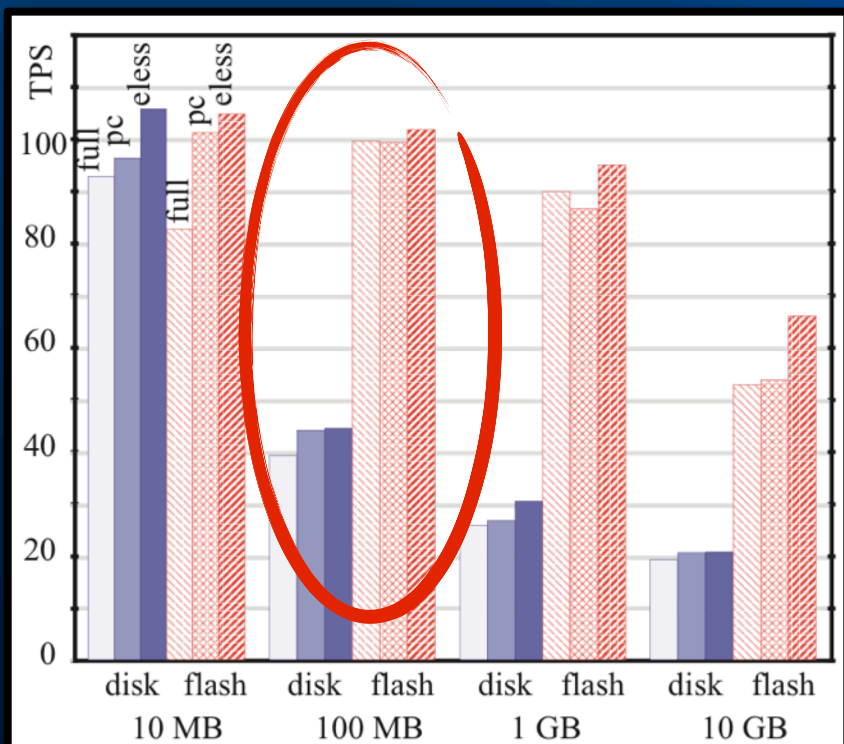


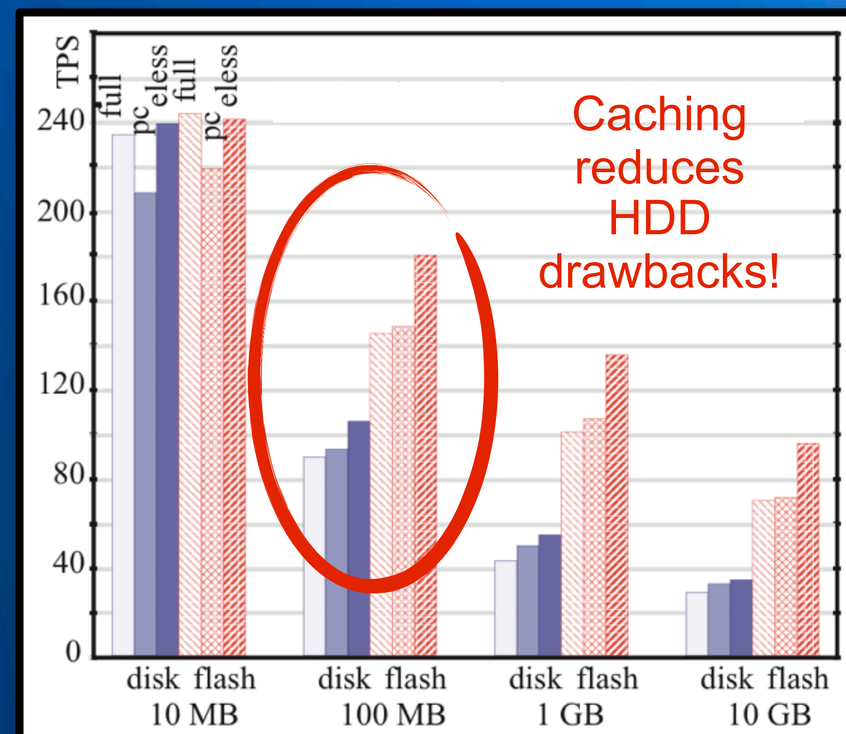


I/O bound

XTC DBMS - Experiments

Read performance: 1000 vs. 10000 Transactions

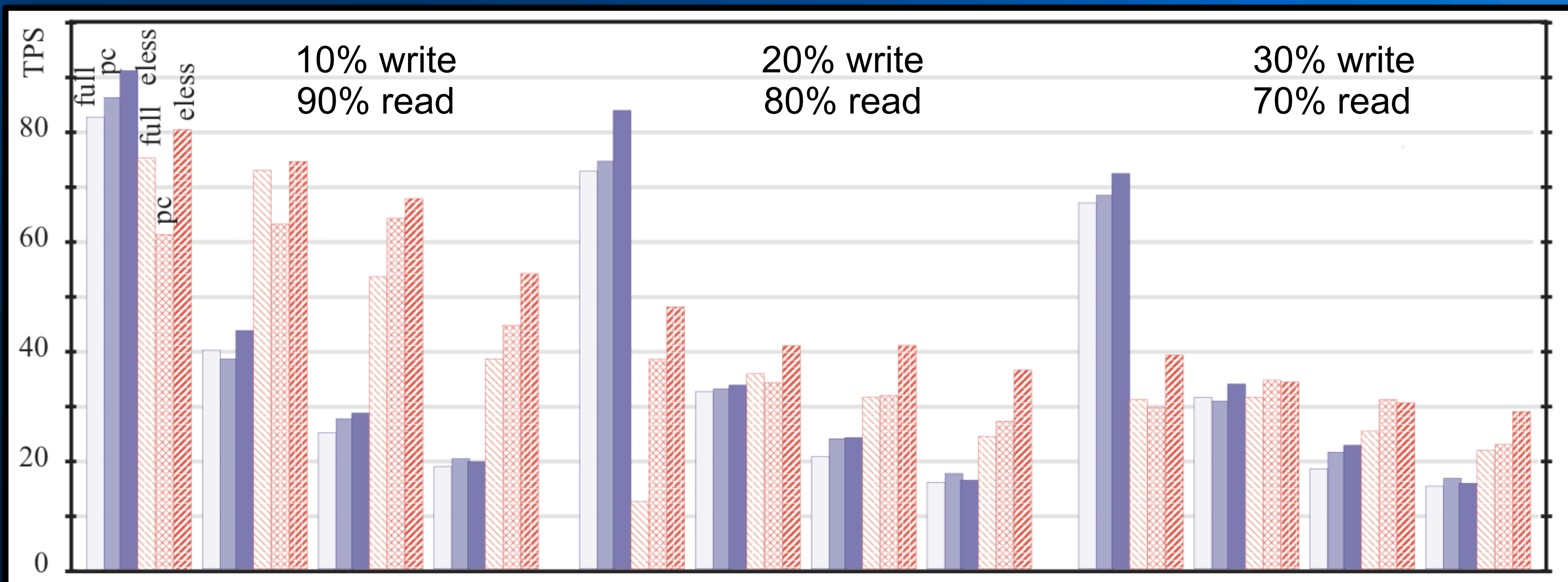




I/O bound

XTC DBMS - Experiments

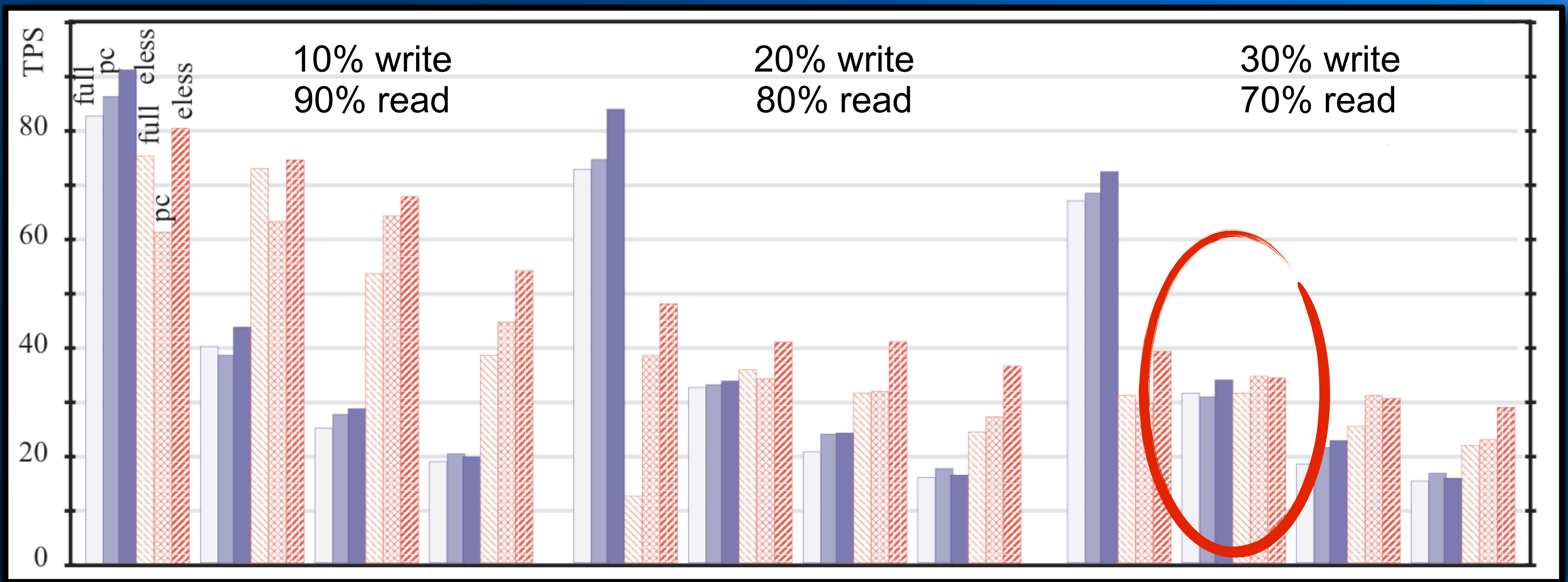
Write performance: 1000 Transactions



I/O bound

XTC DBMS - Experiments

Write performance: 1000 Transactions



I/O bound

SSD vs. HDD - Results

Main Problem:

Papers from 2009/2010

↓

Hardware even older

↓

Results are partially outdated

SSD vs. HDD - Results

Old vs. New:

Device Type	Name	Sequ. Read	Sequ. Write	Random Read	Random Write
Intel X25-M	SSD3	250 MB/s	70 MB/s	35,000 pages/s	3,300 pages/s
Intel 320 Series	NEW	270 MB/s	165 MB/s	39,000 pages/s	21,000 pages/s

SSD vs. HDD - Results

Conclusions:

- Today, high-end SSDs outperform HDDs in all cases.

Performance

Energy Efficiency

SSD vs. HDD - Results

Conclusions:

- Today, high-end SSDs outperform HDDs in all cases.

Performance

Energy Efficiency

Replace all HDDs with SSDs? → Expensive!

SSD vs. HDD - Results

Conclusions:

- **Hybrid Systems:**
  Integrate SSDs into existing environment.
- Storage awareness needed (SSD-aware buffer manager)
- Clever data distribution needed (e.g. Index on SSD).
- Remove unnecessary HDD opmizations (ListPrefetch).
- Exploit efficient Standby-Mode of SSDs.