

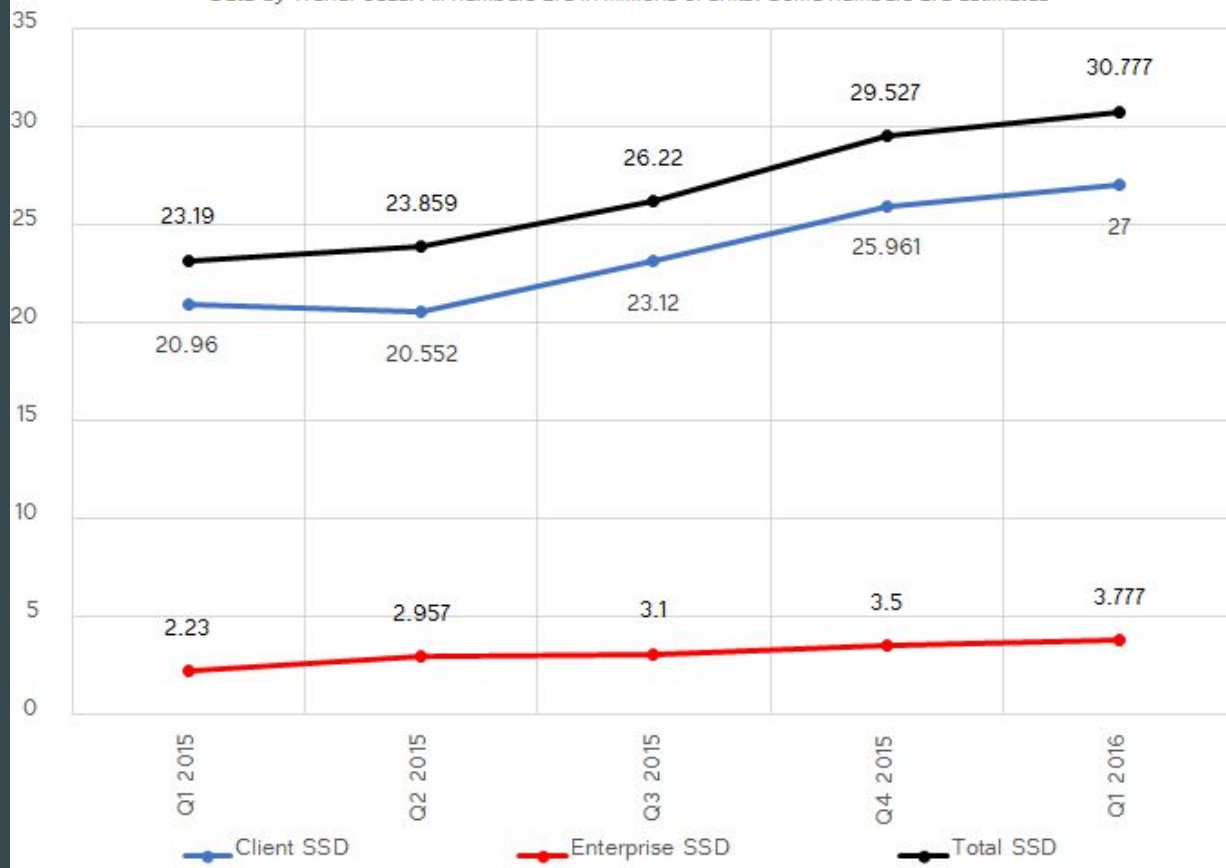
Dyski SSD i ich lepsza integracja z Linuxem poprzez podsystem Open-Channel SSD



Piotr Ciołkosz
18.01.2018

Shipments of SSDs for Different Applications

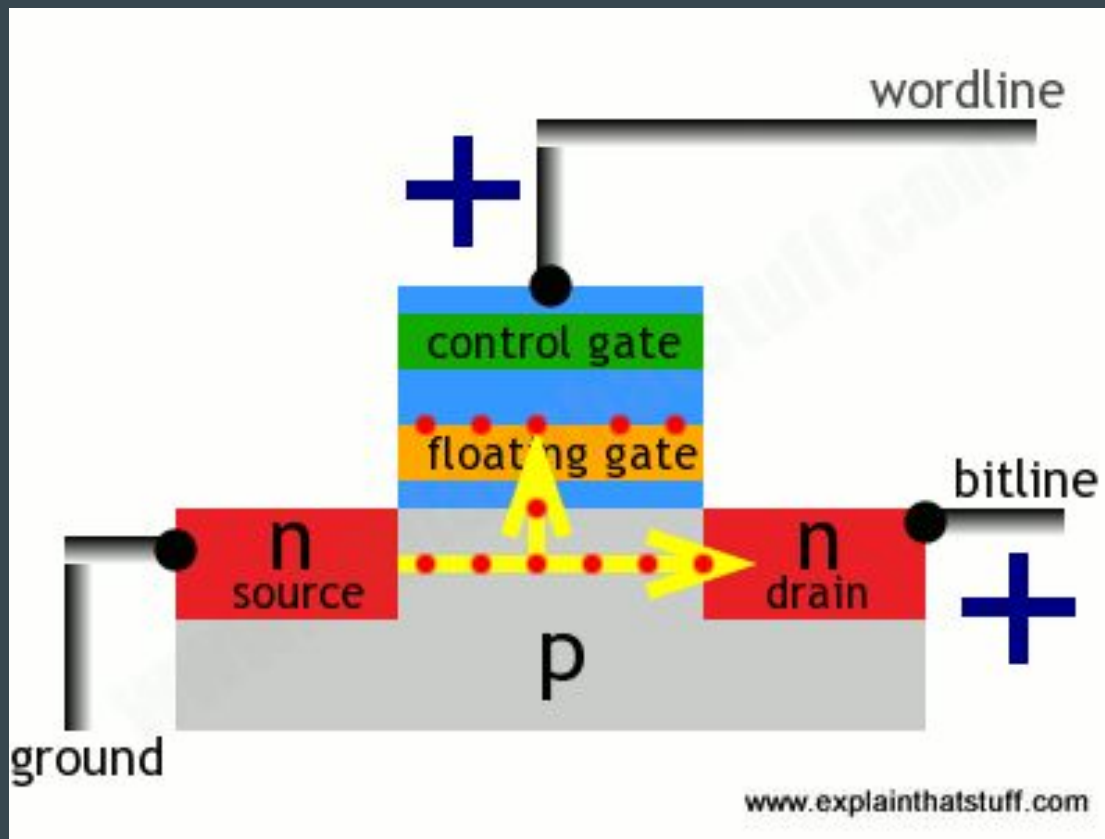
Data by TrendFocus. All numbers are in millions of units. Some numbers are estimates

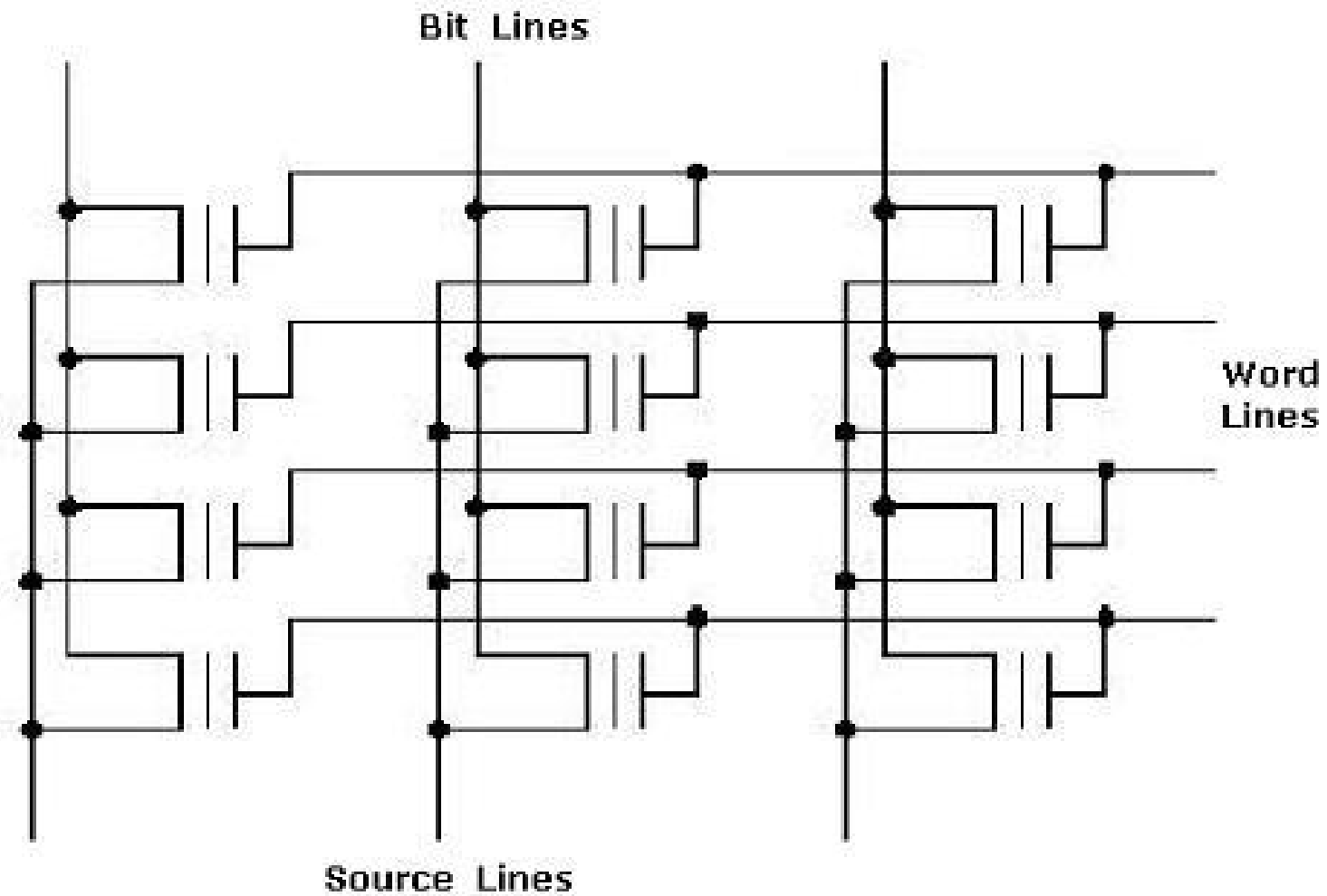


Źródło: <https://www.anandtech.com/show/10348/q1-2016-market-trends-ssds>

Trochę fizyki

- Silne napięcie pomiędzy źródłem i ujściem oraz bramką zamyka elektrony ponad tzw. “floating gate”
- Powodują one przewodnictwo prądu przez umieszczony poniżej półprzewodnik typu “p”
- Reset poprzez przyłożenie jeszcze silniejszego napięcia o przeciwnej polaryzacji





Pamięć flash - budowa wewnętrzna

- Each package contains one or more **dies**.

The die is the smallest unit that can independently execute commands or report status.

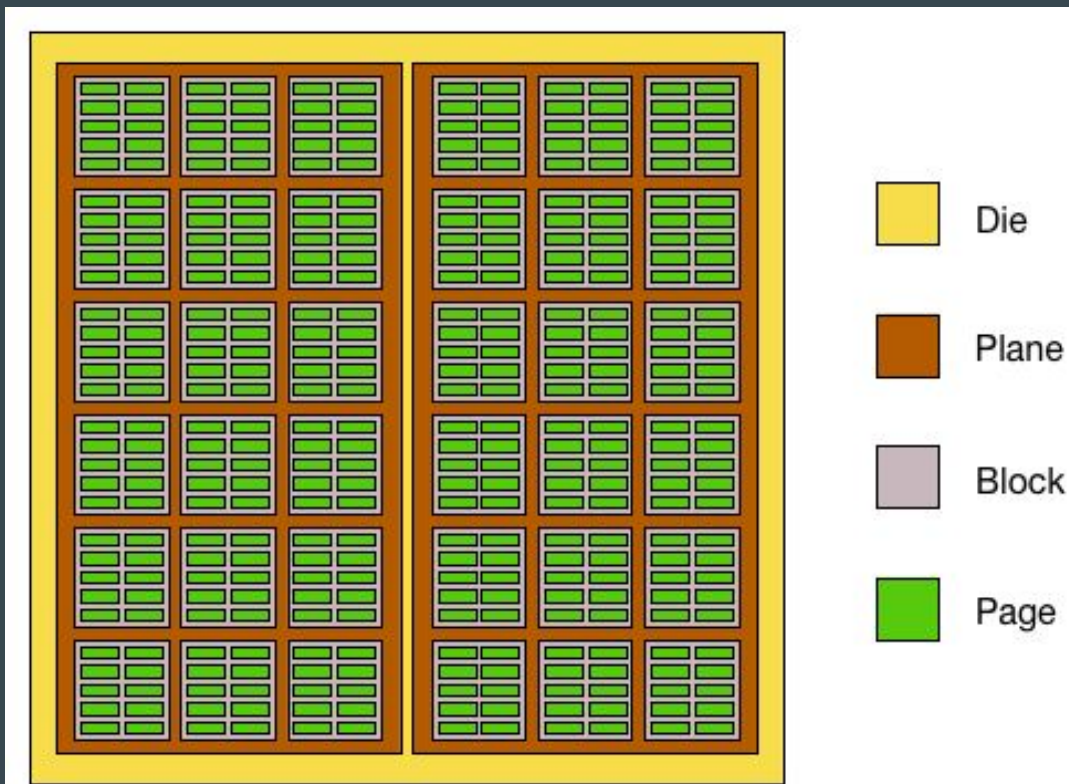
- Each die contains one or more **planes**.

Identical, concurrent operations can take place on each plane, although with some restrictions.

- Each plane contains a number of **blocks**, which are the smallest unit that can be erased.

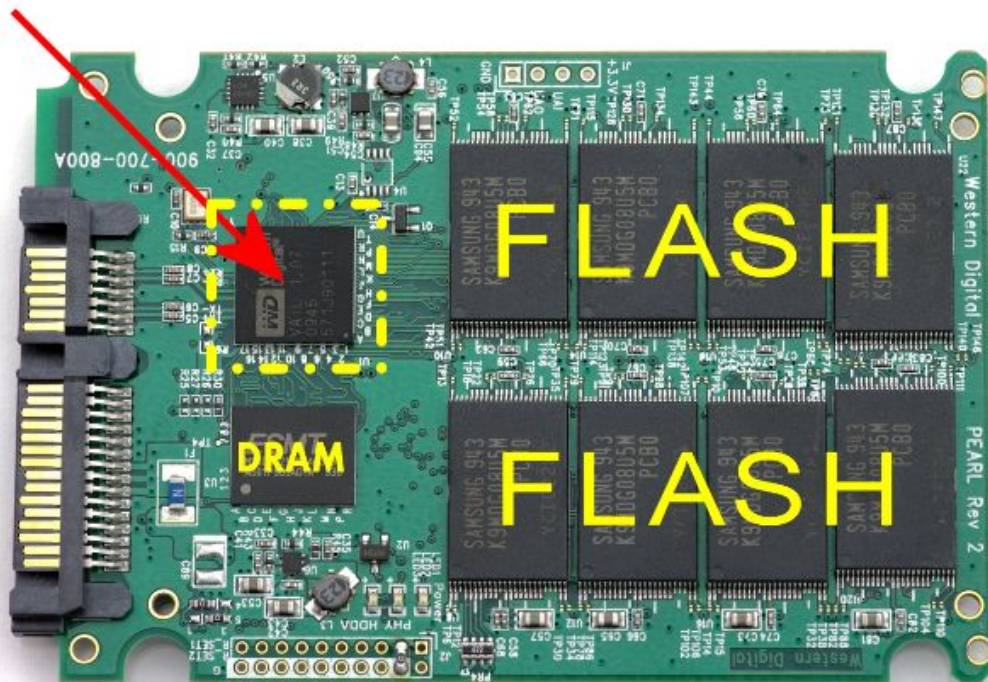
Remember that, it's really important.

- Each block contains a number of **pages**, which are the smallest unit that can be programmed (i.e. written to).



SSD Controller

SATA
and
Power

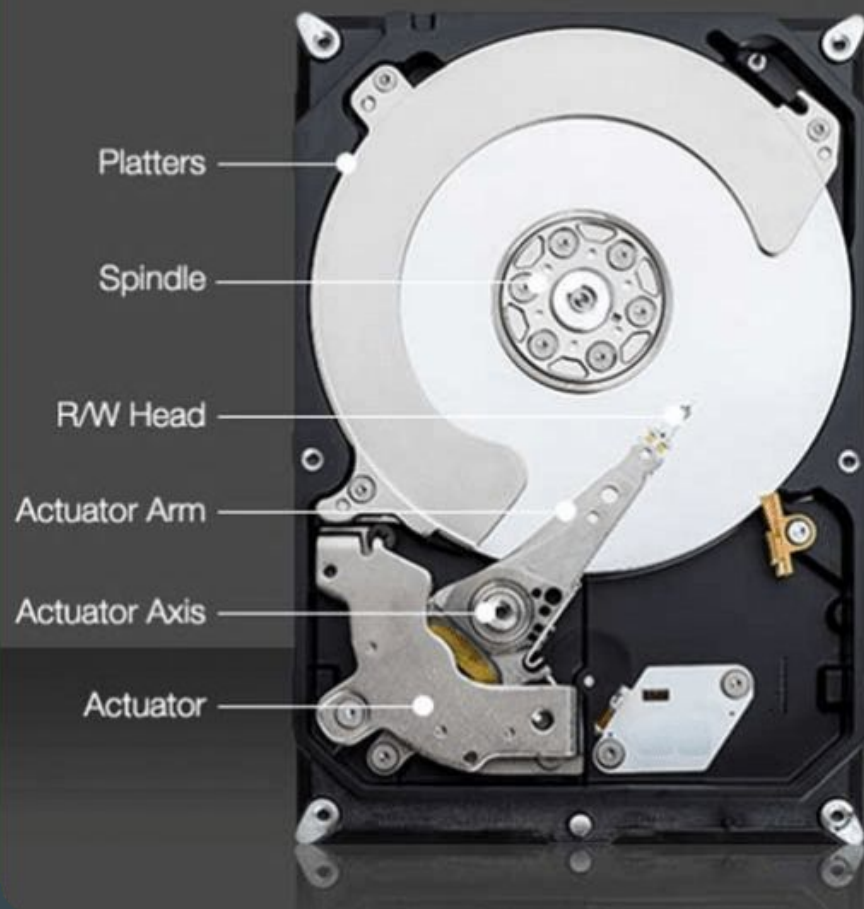


Config and
General I/O

More FLASH
on back

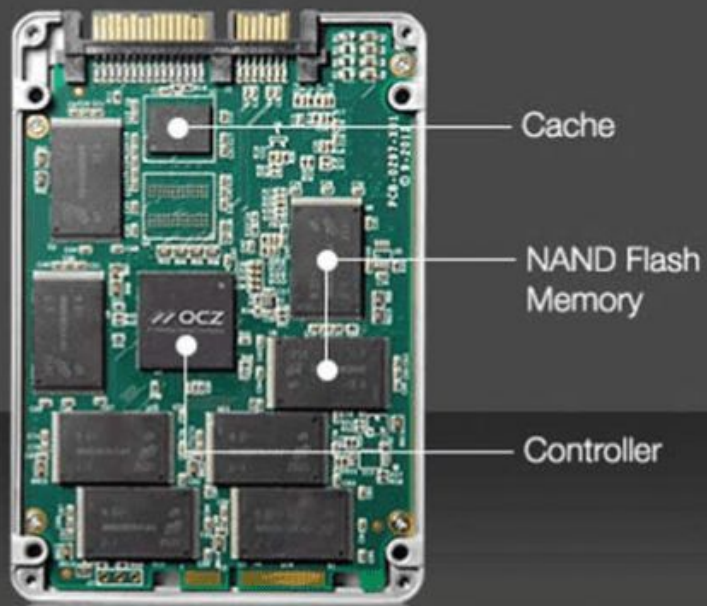
HDD

3.5"



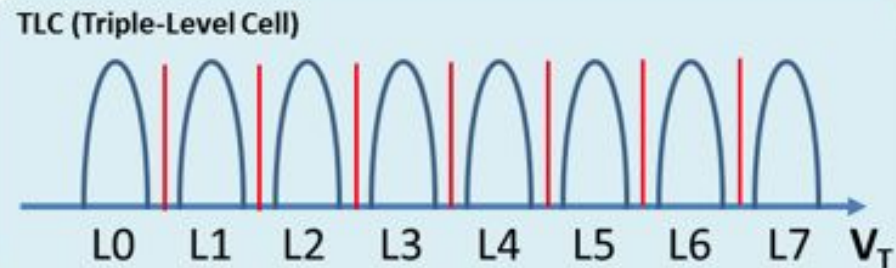
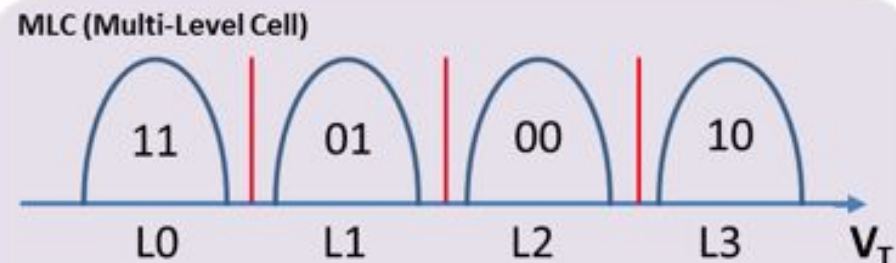
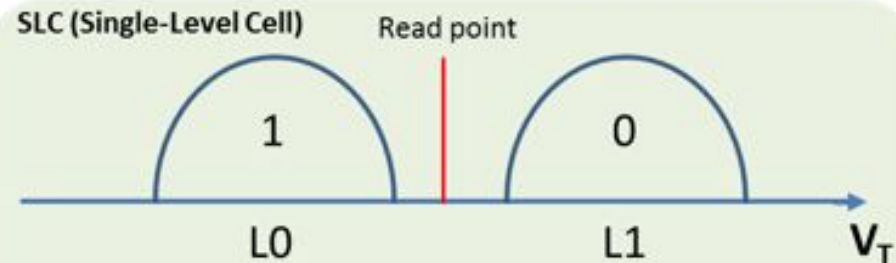
SSD

2.5"



SLC, MLC itp

- Aby zwiększyć pojemność dysków, w jednej komórce można przechowywać więcej niż jeden bit.
- Zamiast napięcie - brak napięcia, rozróżniamy różne poziomy napięcia
- W dyskach konsumenckich zazwyczaj flash jest typu TLC
- W dyskach do profesjonalnych zastosowań, kiedy zależy nam na zmniejszeniu ryzyka utraty danych stosuje się SLC



Dyski SSD jako pamięć do ekstremalnych warunków pracy

- Odporne na przeciążenia aż do setek g przez brak ruchomych części
- Operują w temperaturach do -55°C
- Bardziej odporne na zewnętrzne pole magnetyczne



ilustracja: <https://sklepsamsung.pl/sprzet-it/dyski-ssd/dysk-960-evo-1tb-mz-v6elt0bw>

Dyski SSD - Zalety względem HDD

- Około 15 razy więcej operacji na sekundę
- Stukrotnie mniejsza latencja odczytu danych
- Mniejszy rozmiar
- Cicha praca
- Mniejsze zużycie prądu



Wady i problemy

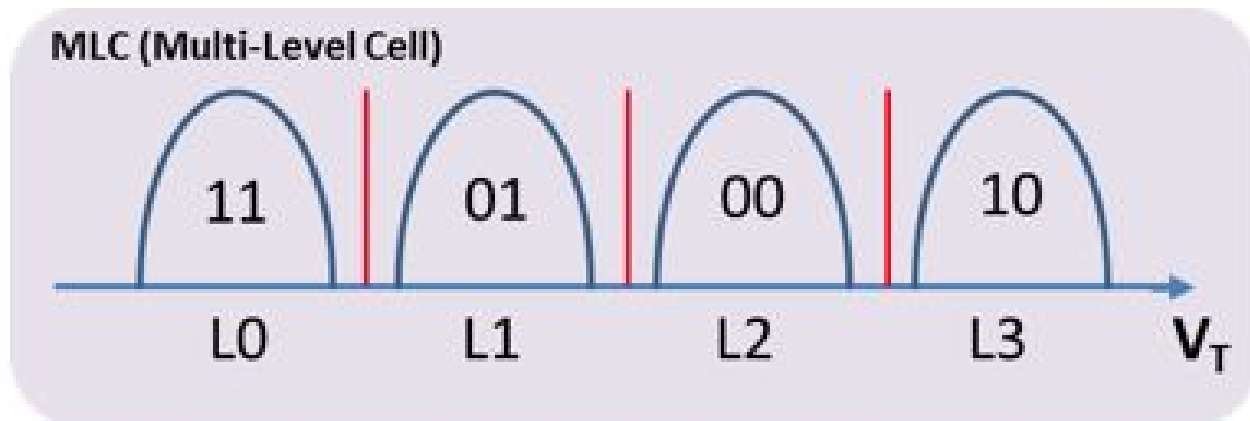
Pomimo zdecydowanej przewagi w szybkości działania, SSD ma też szereg wad.

Mało ciekawa i oczywista wada:

Cena, pomimo jej ciągłego spadku



Oczekiwania...



Rzeczywistość...

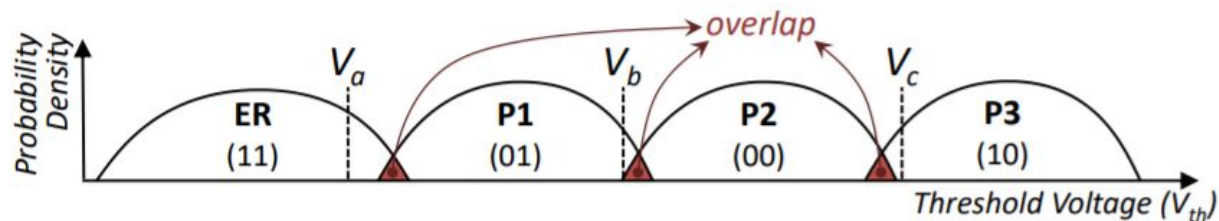


Figure 13: Threshold voltage distribution shifts and widening can cause the distributions of two neighboring states to overlap with each other (compare to Figure 7), leading to read errors.

Wada #1: Floating gate się zużywa

Każda komórka pamięci ma ograniczoną oczekiwaną liczbę zapisów, którą można na niej wykonać.

Przyczyną tego, są defekty floating gate wywoływane przez przykładanie do nich dużego napięcia.

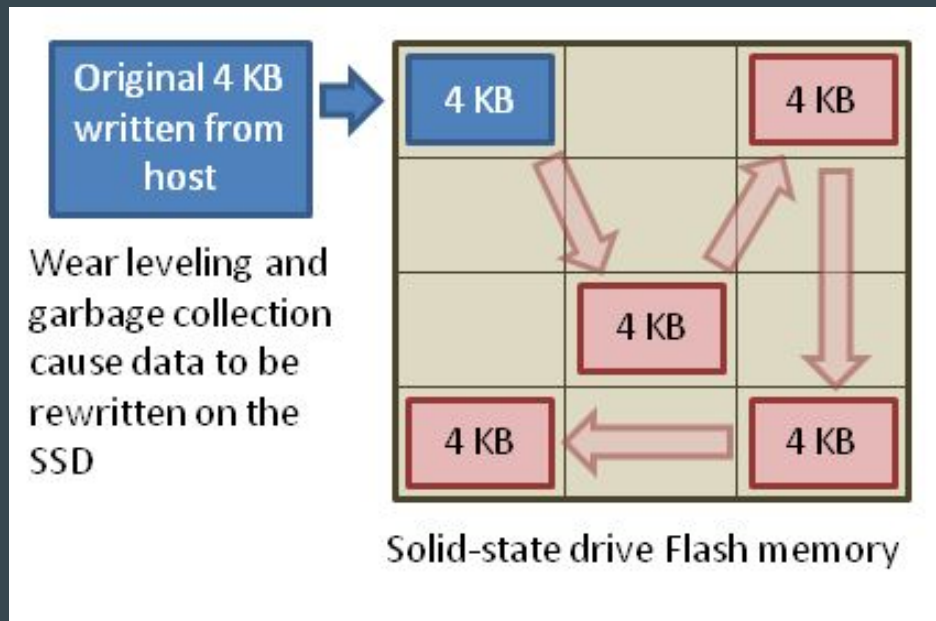
Z czasem elektrony mogą zacząć przez nie przeciekać, co powoduje błędy zapisu.

Komórka błędnie zapisana jest przez kontroler oznaczana jako zepsuta i przestaje być używana.

Wada #1: Floating gate się zużywa

Aby zminimalizować ten efekt, kontrolery zliczają zapisy do konkretnych stron i starają się na tej podstawie balansować przyszłe zapisy.

Dzięki temu opóźniają moment pierwszych błędów zapisy przez zużycie komórek.



Wada #1: Floating gate się zużywa

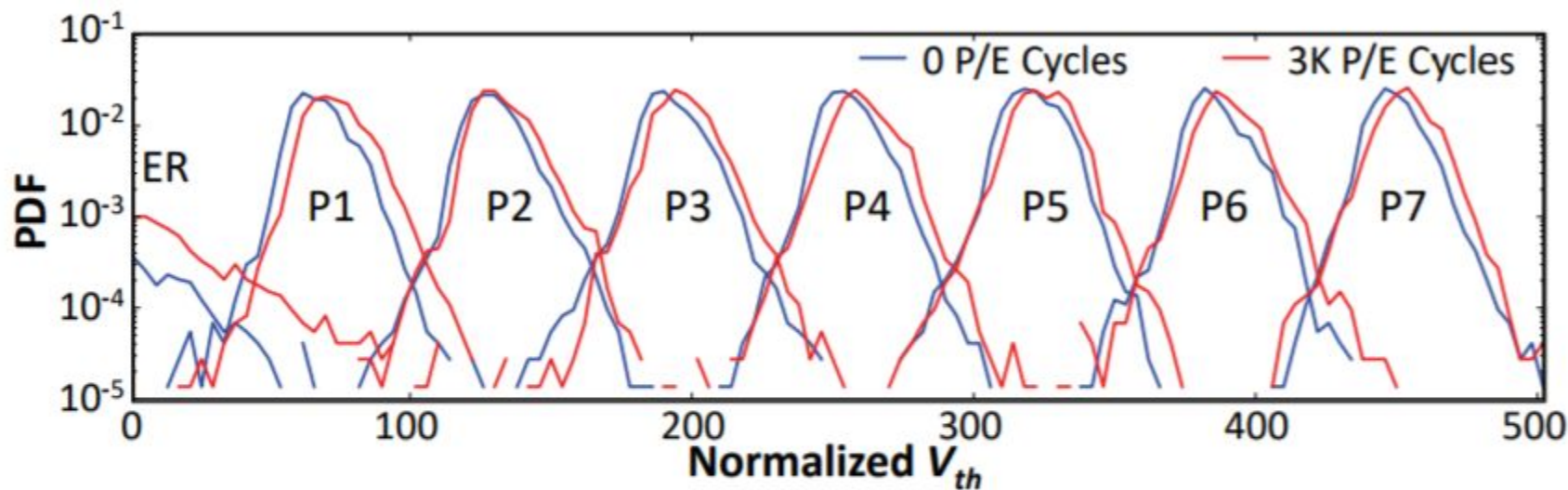


Figure 14: Threshold voltage distribution of TLC NAND flash memory after 0 P/E cycles and 3,000 P/E cycles.

Wada #2.1: Zapis wymaga usunięcia zawartości komórki, a to długo trwa.

Zapisana strona nie może być zapisana, bez uprzedniego usunięcia jej zawartości.

Operacja usunięcia trwa bardzo długo - nawet kilka milisekund.

Aby ukryć latencję zapisu kontrolery starają się wykonać zapis do nowej komórki pamięci i przemapować adres na tą nową komórkę, a reset zawartości starej zostawić sobie na później.

Dlatego właśnie dobrze jest nie wypełniać pamięci flash do pełna, bo mocno zwiększa to latencję zapisu.

Wada #2.2: Operacja usunięcia może być wykonana tylko na dużych blokach

Usunięcie zawartości komórki wymaga sporego napięcia.

Z tego powodu pamięci flash pozwalają na usuwanie tylko całych bloków, często jest to kilka MB.

Block X	A	B	C
	D	free	free
	free	free	free
	free	free	free

Block Y	free	free	free
	free	free	free
	free	free	free
	free	free	free

1. Four pages (A-D) are written to a block (X). Individual pages can be written at any time if they are currently free (erased).

Block X	A	B	C
	D	E	F
	G	H	A'
	B'	C'	D'

Block Y	free	free	free
	free	free	free
	free	free	free
	free	free	free

2. Four new pages (E-H) and four replacement pages (A'-D') are written to the block (X). The original A-D pages are now invalid (stale) data, but cannot be overwritten until the whole block is erased.

Block X	free	free	free
	free	free	free
	free	free	free
	free	free	free

Block Y	free	free	free
	free	E	F
	G	H	A'
	B'	C'	D'

3. In order to write to the pages with stale data (A-D) all good pages (E-H & A'-D') are read and written to a new block (Y) then the old block (X) is erased. This last step is *garbage collection*.

Wada #3: Dane w pamięci flash mogą z czasem zanikać

Komórka pamięci nie jest w 100% szczelna. Dodatkowo odczyty z komórki mogą wpływać na ładunek w niej przechowywany.

Elektrony przechowywane w floating gate mogą z czasem uciekać, przez co dane muszą być periodycznie czytane i ponownie zapisywane, aby uniknąć ich przekłamania.

Podaje się czas rzędu ~ 2 lat, dlatego SSD słabo nadają się do archiwizacji.

Wada #3: Dane w pamięci flash mogą z czasem zanikać

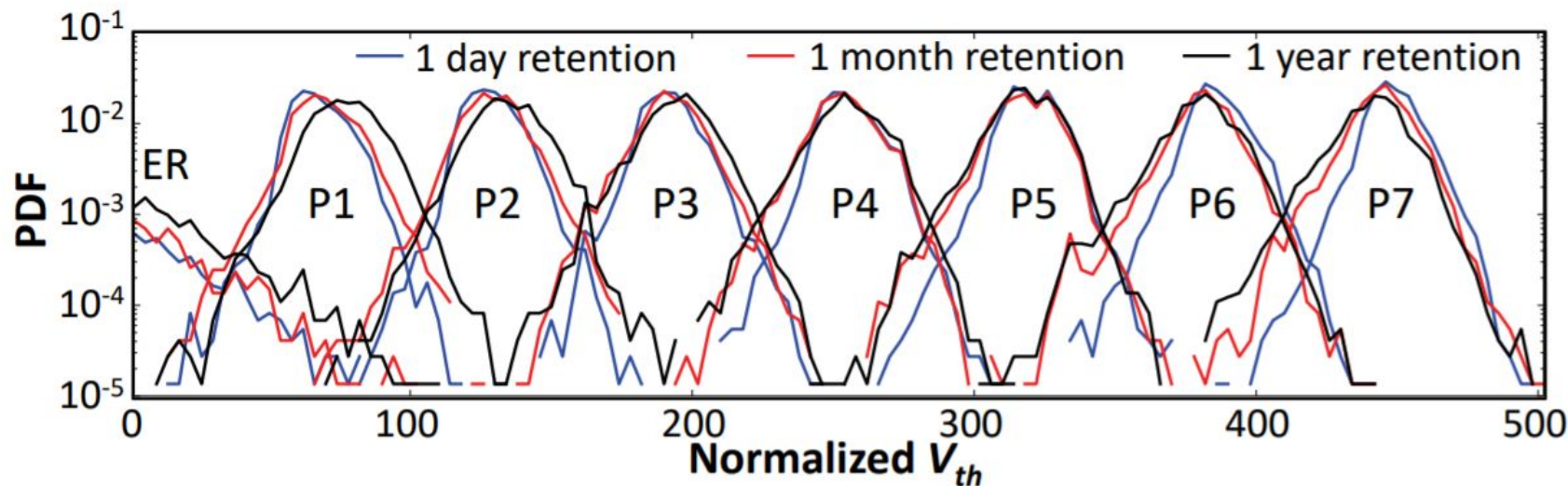


Figure 18: Threshold voltage distribution for TLC NAND flash memory after one day, one month, and one year of retention time.

Wada #3: Dane w pamięci flash mogą z czasem zanikać

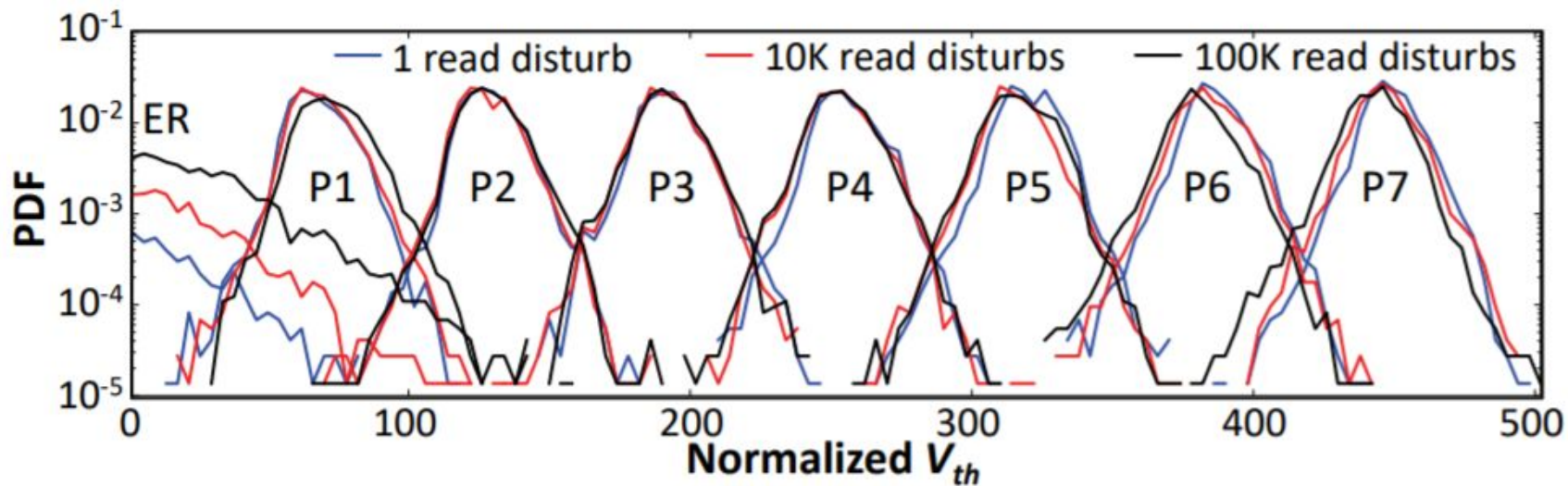
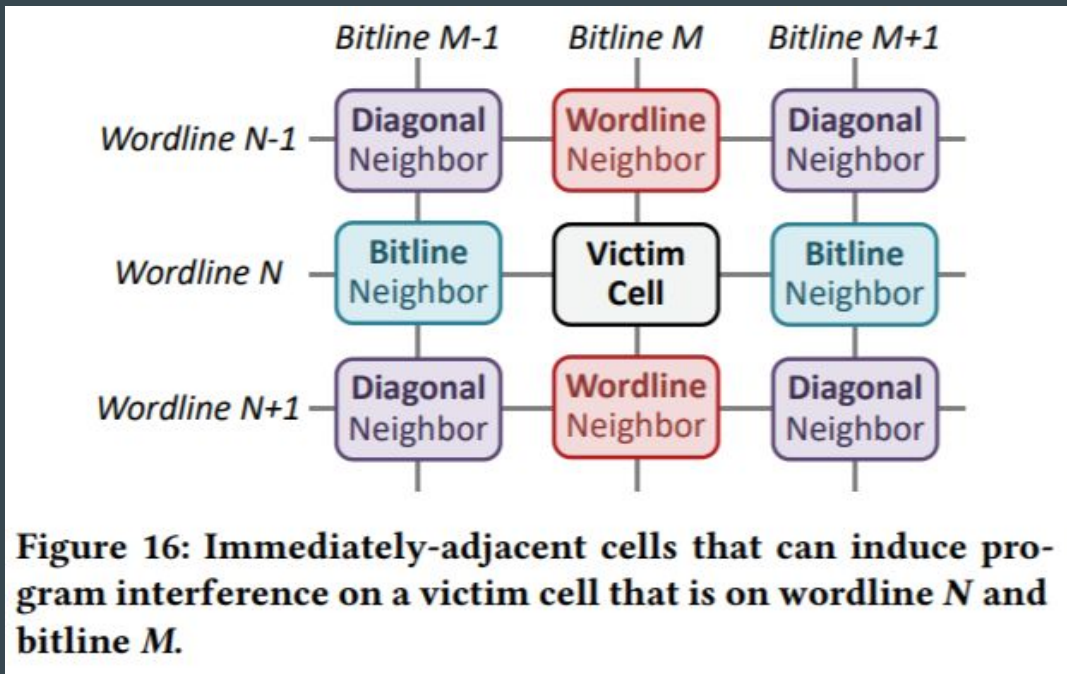


Figure 19: Threshold voltage distribution for TLC NAND flash memory after 1, 10K, and 100K read disturb operations.

Wada #4: Interferencja między komórkami

Przez potrzebę zastosowania dużych napięć przy zapisie i resecie oraz coraz gęstszemu upakowaniu komórek pamięci ciężko jest w 100% uodpornić komórki na działania na ich sąsiadach.

Każda operacja na komórce może wpłynąć na ładunek w sąsiedniej komórce.



Wada #5: Niejednostajna latencja odczytu

Z powodu wad związanych z przekłamaniami zawartości komórek, pamięci flash trzymają dodatkowe dane używane do korekcji błędów.

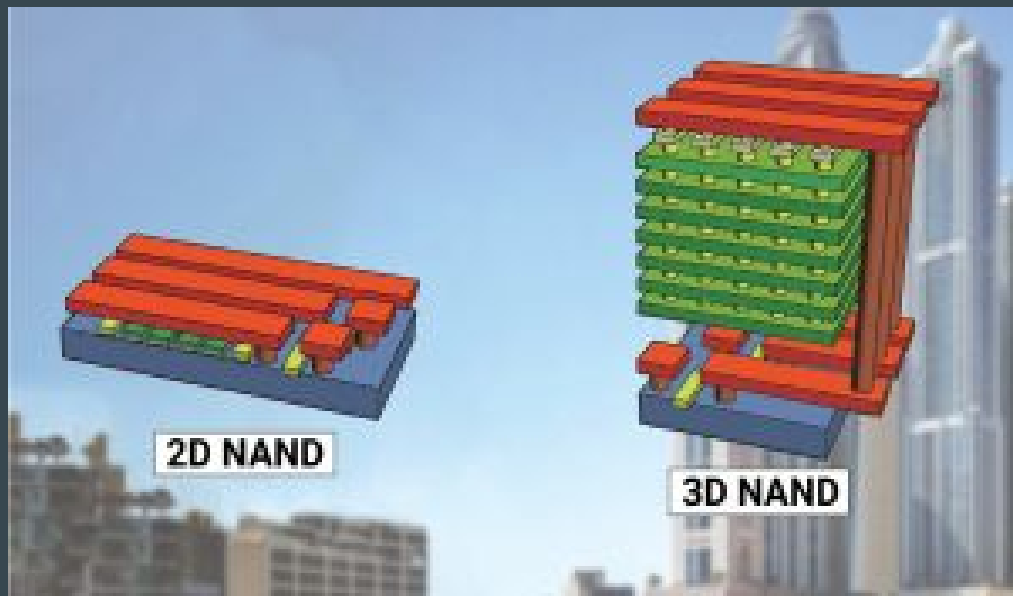
Po odczycie sprawdzana jest poprawność, w przypadku błędu następuje próba naprawy danych, w przypadku porażki odczyt jest ponawiany (czasem wielokrotnie) z innym napięciem.

Dodatkowo odczyty muszą czekać, aż zapis bloku się zakończy, co może trwać długo jeśli zapis wiązał się z usunięciem poprzedniego bloku.

Powoduje to zmienną latencję odczytu.

3D NAND flash

- Warstwy ułożone jedna na drugiej, co zwiększa upakowanie pamięci.
- Inny materiał użyty w floating gate, bardziej odporny na zużycie, ale też bardziej podatny na ucieczkę ładunku.
- Większy feature scaling (50 nm vs 15-20 nm), dzięki czemu mniejszy wpływ komórek na sąsiadów, jednak więcej każda komórka ma więcej interferujących sąsiadów.



TRIM command

HDD nie potrzebowało wiedzieć kiedy ktoś zwalnia daną pamięć, więc w protokole komunikacji nie było odpowiedniego komunikatu.

SSD lubi mieć dużo wolnych obszarów do usunięcia ich zawartości w wolnej chwili dla przyspieszenia zapisu i mniejszej fragmentacji.

	1. User writes four new files	2. User deletes file "C"	3. User writes new file "E"
OS Logical View			
SSD <u>Logical</u> View (LBAs)			
SSD <u>Physical</u> View			
	SSD writes new data; only SSD knows about OP	Only OS knows location C1 & C2 are no longer valid and SSD keeps rewriting it during GC	OS writes new file to old location; SSD marks old location ready for GC and file E gets written elsewhere

ilustracja: <http://www.thessdreview.com/daily-news/latest-buzz/garbage-collection-and-trim-in-ssds-explained-an-ssd-primer/2/>

NVMe

Interface stworzony z myślą o dyskach SSD, aby zastąpić ten projektowany dla HDD np. SATA. Zaprojektowany by lepiej wykorzystać równoległość. Wprowadzony około 2011/2012 roku.

High-level comparison of AHCI and NVMe ^[5]		
	AHCI	NVMe
Maximum queue depth	One command queue; 32 commands per queue	65535 queues; ^[29] 65536 commands per queue
Uncacheable register accesses (2000 cycles each)	Six per non-queued command; nine per queued command	Two per command
<u>MSI-X</u> and interrupt steering	A single interrupt; no steering	2048 MSI-X interrupts
Parallelism and multiple threads	Requires synchronization lock to issue a command	No locking
Efficiency for 4 KB commands	Command parameters require two serialized host DRAM fetches	Gets command parameters in one 64-byte fetch

Doświadczenia z SSD w Google.

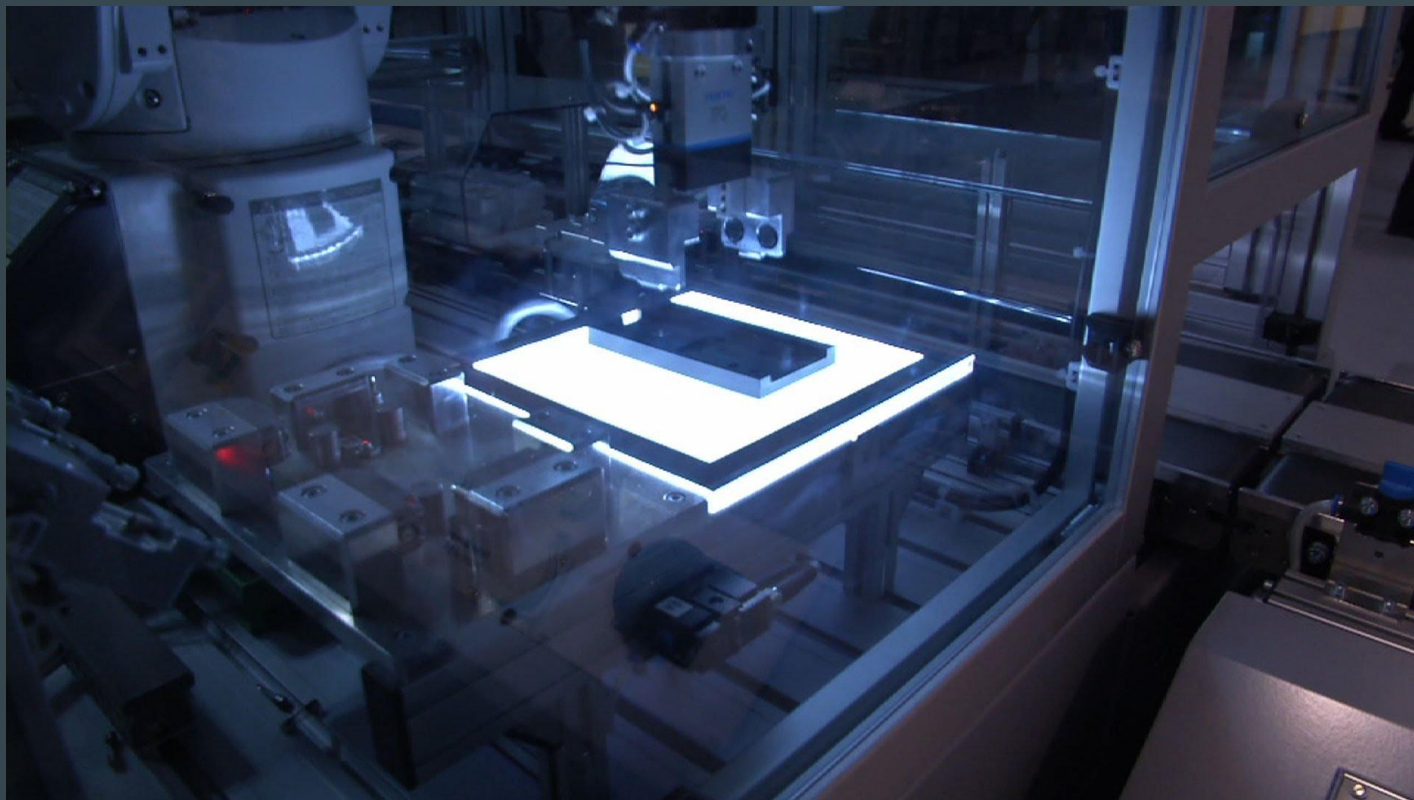
Opracowanie danych zebranych podczas używania 10 różnych dysków MLC i SLC przez 6 lat w centrach danych Google'a

Wyciągnięte wnioski:

- 20-63% dysków doświadcza nieodwracalnego błędu odczytu w ciągu 4 lat pracy.
- 4-10% dysków musiało zostać wymienionych w ciągu 4 lat pracy.
- Liczba błędów rośnie wraz z cyklami zapisu, jednak nie tak szybko jak sądzono.
- Fizyczny wiek dysku też ma znaczenie na jego niezawodność.
- SLC nie jest bardziej niezawodne niż MLC.
- Zazwyczaj dyski mają kilka lub bardzo dużo zepsutych bloków.
- Dla porównania, 2-9% dysków HDD jest wymienianych ROCZNIE, ale tylko 3.5% z nich doświadcza zepsutych bloków w ciągu 3 lat pracy.

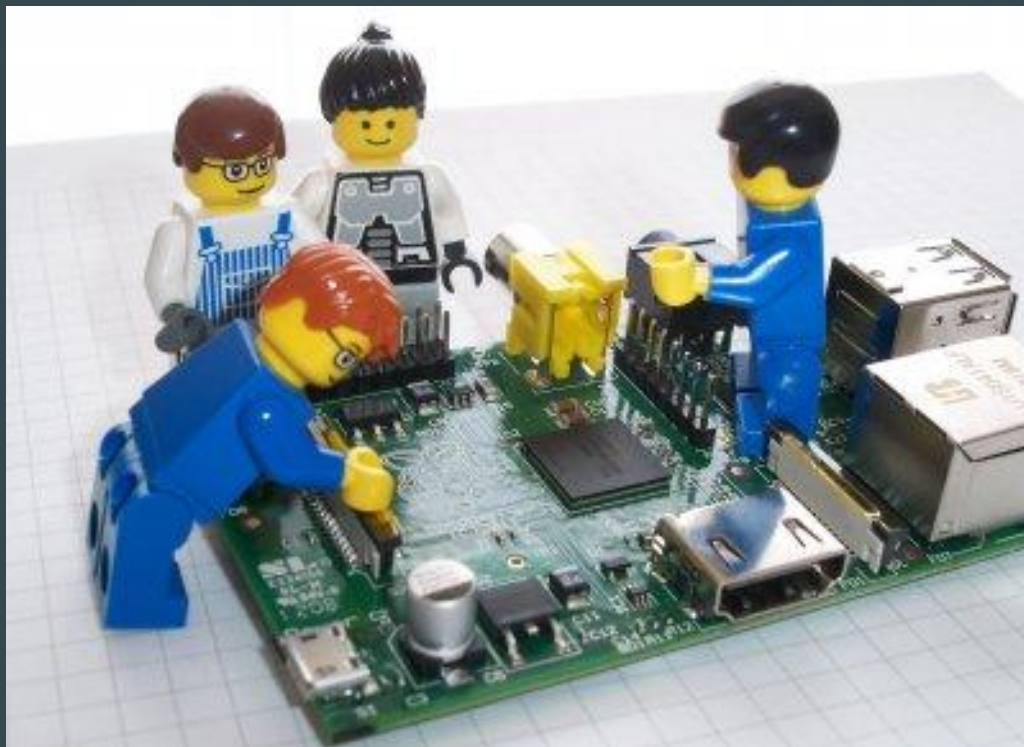
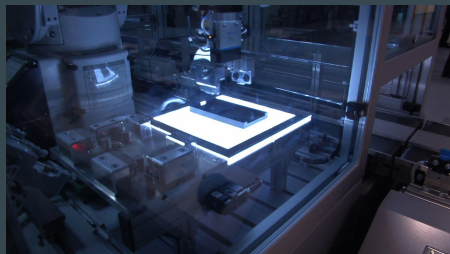
Co zrobić aby było lepiej?

Co zrobić aby było lepiej?



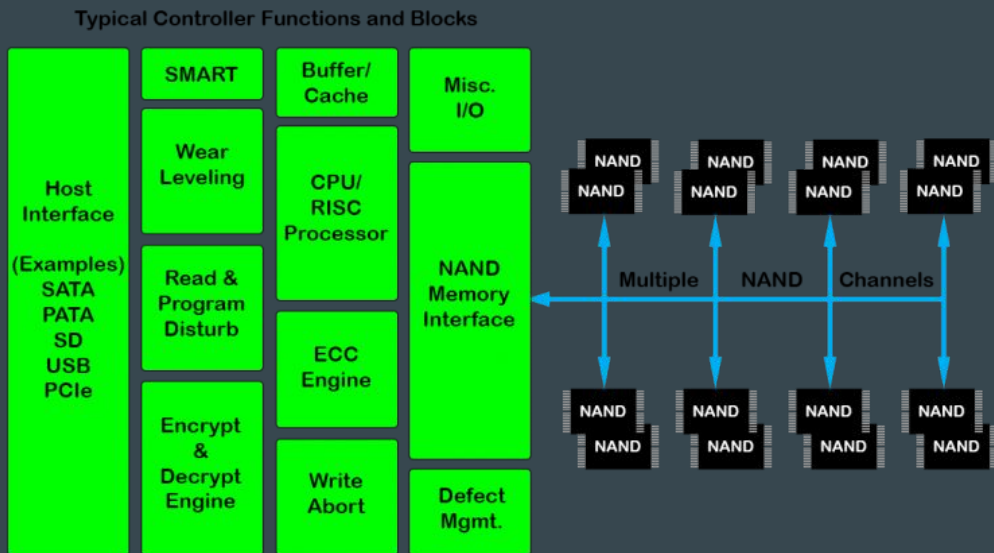
ilustracja: <https://www.youtube.com/watch?v=bvyS2YSmAsE>

Co zrobić aby było lepiej?



Kontroler

- Translacja adresów logicznych na fizyczne
- Buforowanie zapisów
- Error correction and handling
- Managing parallelism
- Level-wearing
- Garbage Collecting



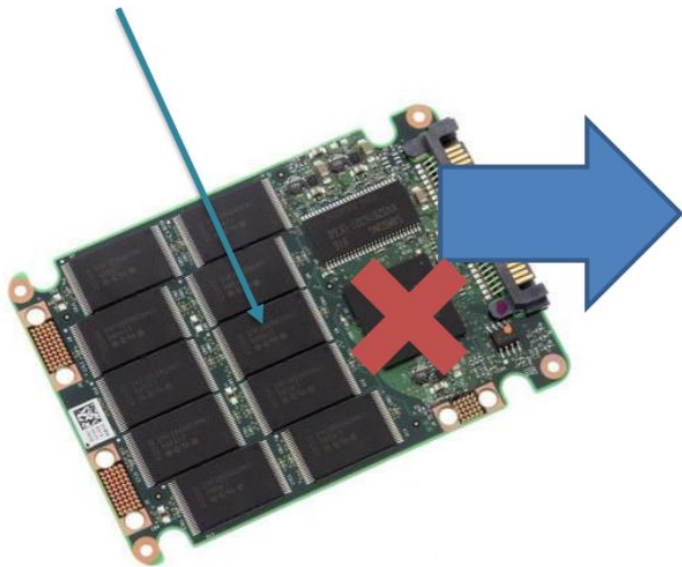
Generic Solid State Drive (SSD) Controller Architecture

Pomysł: Uprościć kontroler przy przerzuceniu części odpowiedzialności na hosta

- Możliwość dostosowania poszczególnych trade-offów do danej aplikacji
 - Większa swoboda implementowania algorytmów rozmieszczania danych w zależności od aplikacji
 - Umożliwienie aplikacjom wgląd w rozmieszczenie danych i podejmowanie decyzji na tej podstawie.
-
- Wymaga dodatkowych (drobnych) zasobów komputera

Open-Channel SSDs

Physical flash exposed to the host (Read, write, erase)



Host

- Data placement
- IO Scheduling
- Over-provisioning
- Garbage collection
- Wear levelling

LightNVM: Podsystem Open-channel SSD dla linuxa

- 3 różne interface'y w zależności jak dokładnie chcemy kontrolować dysk.

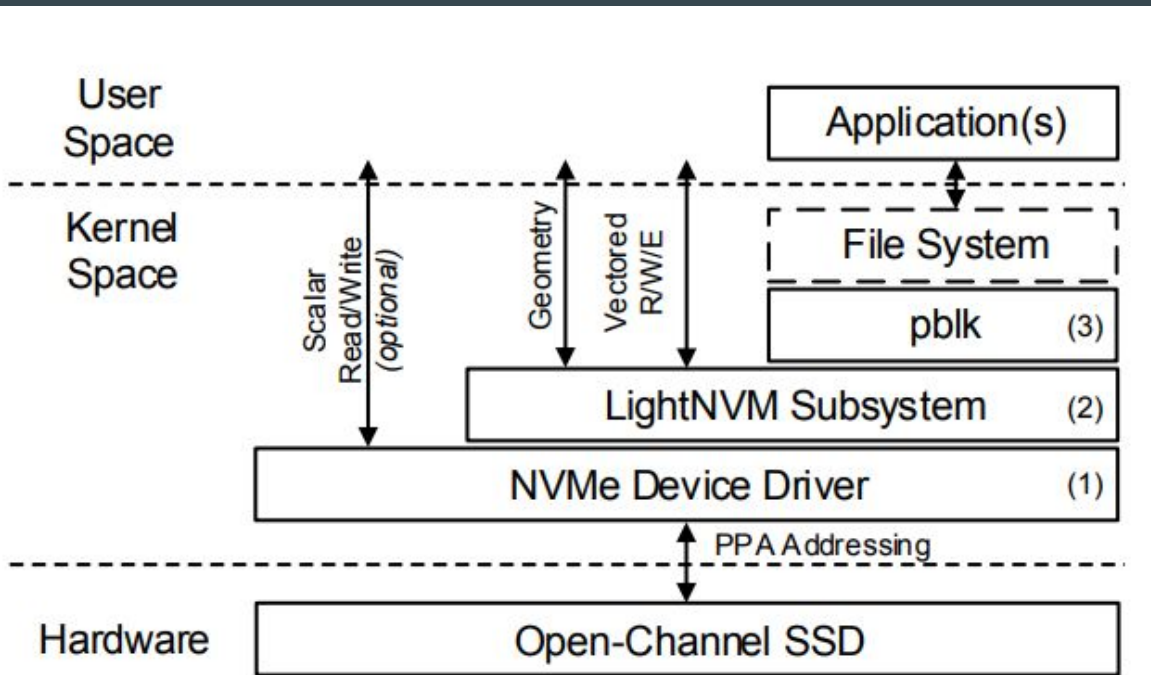
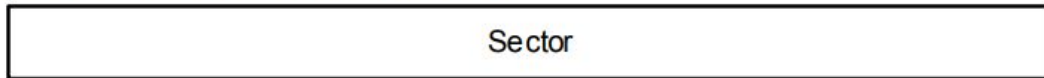


Figure 3: LightNVM Subsystem Architecture

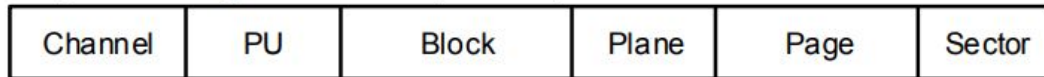
PPA

- Nowe adresowanie świadome możliwości równoległej pracy SSD nazwane PPA (Physical Page Address).
- Sterownik dysku dostosowany do użycia z LightNVM i komunikacji przy użyciu PPA.

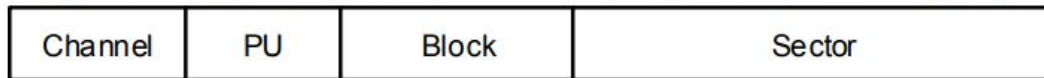
Logical Block Address (LBA)



Physical Page Address (NAND)



Physical Page Address (NAND with Sector-sized Pages)



MSB

64bit

LSB

Figure 2: Logical Block Addresses compared to Physical Page Addresses for NAND flash.

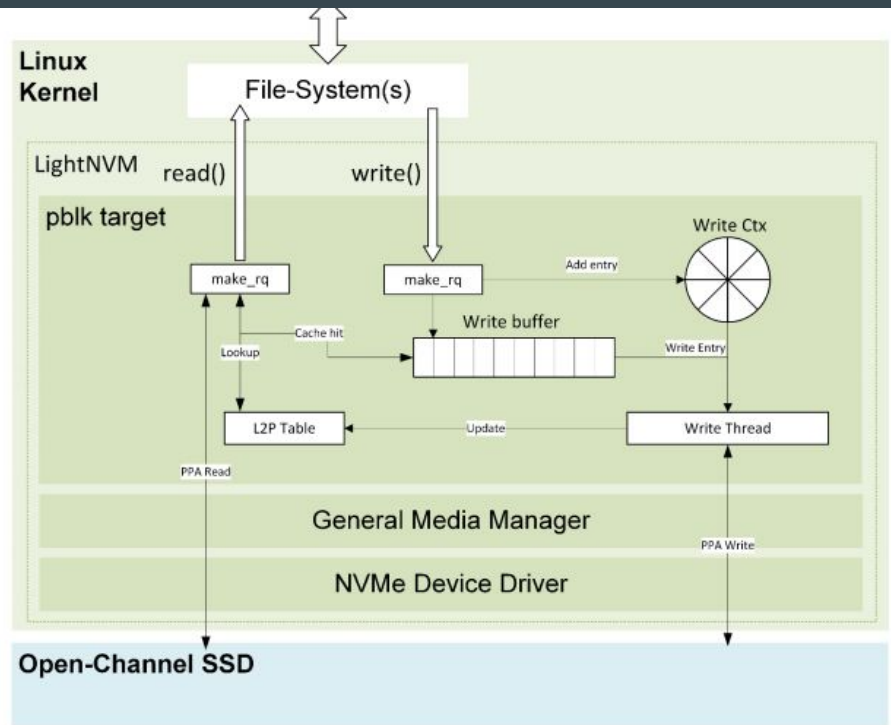
Podsystem LightNVM

- Pozwala dowiedzieć się o geometrii urządzenia (Ile kanałów, PU, bloków itp). Dzięki temu użytkownik może zaprojektować własny FTL, który rozłoży dane na dysku w sposób zależny od konkretnego zastosowania.
- Zwektoryzowane polecenia z oddzielnym polem statusu dla każdego adresu. Zapewnia to większą przepustowość oraz kontrolę błędów.

pblk

pblk – Physical Block Target (FTL)

- Exposes media as a LBA block device
- Implements write buffering
- Manages logical to physical address table (L2P)
- Orchestrates data placement
- Garbage collection
- Multiple targets per device



Wyniki

Test dwóch wątków, jeden zapisuje dane, inny czyta rozłączne dane.

Dzięki umiejętnemu rozmieszczeniu danych udało się zmniejszyć latencje odczytu, a co za tym idzie, również latencje zapisu.

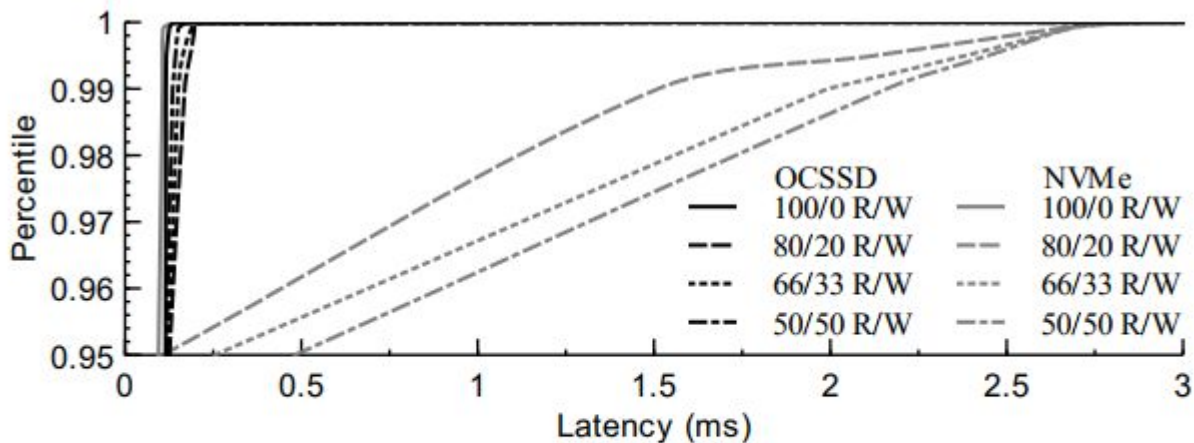


Figure 8: Latency comparison of OCSSD and NVMe SSD showing how writes impact read latencies. Note that the y-axis does not start at 0, but at 0.95.

Wyniki 2

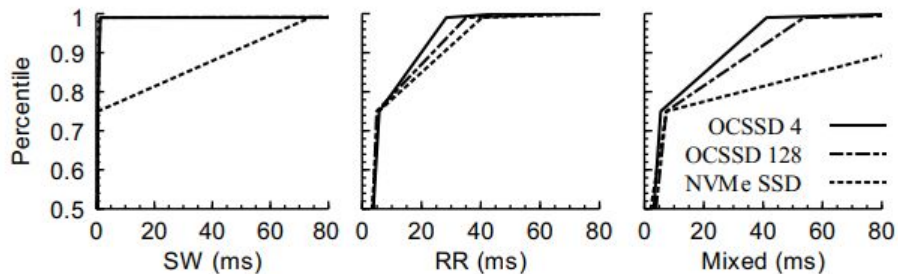


Figure 6: Latencies for RocksDB sequential writes, random reads and mixed workloads on OCSSD and NVMe SSD

	NVMe SSD	OCSSD 128	OCSSD 4
SW	276	396	80
RR	5064	5819	5319
Mixed	2208	3897	4825

Table 2: Throughput (MB/s) for RocksDB sequential writes, random reads, and mixed workloads on OCSSD and NVMe SSD

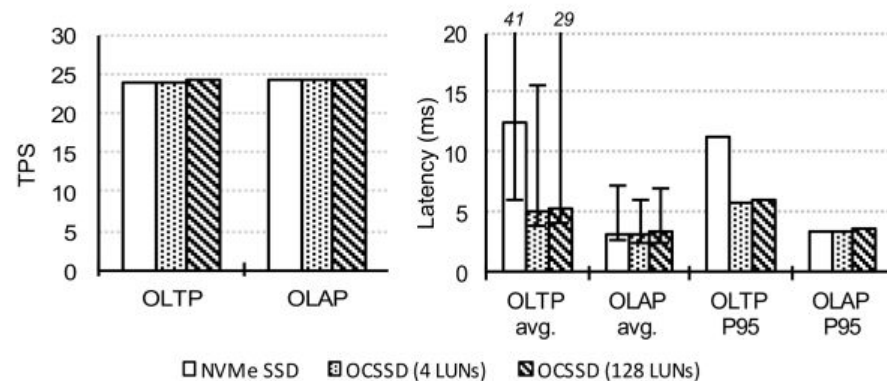


Figure 7: Transactions per second and latencies for OLTP and OLAP on NVMe SSD and OCSSD.

LightNVM in kernel 4.4

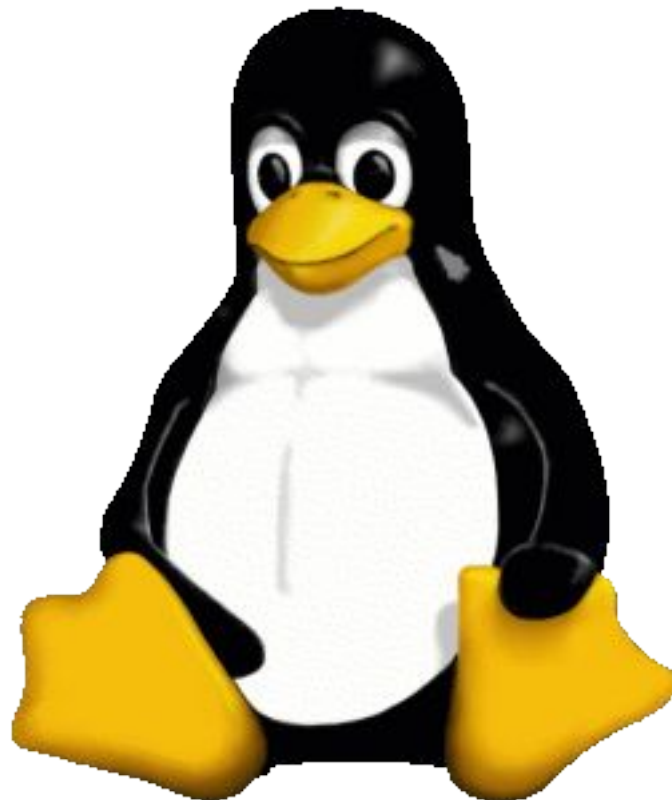
Filter tags

/ include / linux / lightnvm.h

Search Identifier

```
v4.4.30
v4.4.29
v4.4.28
v4.4.27
v4.4.26
v4.4.25
v4.4.24
v4.4.23
v4.4.22
v4.4.21
v4.4.20
v4.4.19
v4.4.18
v4.4.17
v4.4.16
v4.4.15
v4.4.14
v4.4.13
v4.4.12
v4.4.11
v4.4.10
v4.4.9
v4.4.8
v4.4.7
v4.4.6
v4.4.5
v4.4.4
v4.4.3
v4.4.2
v4.4.1
v4.4
v4.4-rc8
v4.4-rc7
v4.4-rc6
v4.4-rc5
v4.4-rc4
v4.4-rc3
v4.4-rc2
v4.4-rc1
v4.3
v4.2
v4.1
v4.0
v3
v2.6
```

```
118 struct nvm_tgt_instance {
119     struct nvm_tgt_type *tt;
120 };
121
122 #define ADDR_EMPTY (~0ULL)
123
124 #define NVM_VERSION_MAJOR 1
125 #define NVM_VERSION_MINOR 0
126 #define NVM_VERSION_PATCH 0
127
128 #define NVM_BLK_BITS (16)
129 #define NVM_PG_BITS (16)
130 #define NVM_SEC_BITS (8)
131 #define NVM_PL_BITS (8)
132 #define NVM_LUN_BITS (8)
133 #define NVM_CH_BITS (8)
134
135 struct ppa_addr {
136     /* Generic structure for all addresses */
137     union {
138         struct {
139             u64 blk      : NVM_BLK_BITS;
140             u64 pg       : NVM_PG_BITS;
141             u64 sec       : NVM_SEC_BITS;
142             u64 pl        : NVM_PL_BITS;
143             u64 lun       : NVM_LUN_BITS;
144             u64 ch        : NVM_CH_BITS;
145         } g;
146
147         u64 ppa;
148     };
149 };
150
151 struct nvm_rq {
152     struct nvm_tgt_instance *ins;
153     struct nvm_dev *dev;
154
155     struct bio *bio;
156
157     union {
158         struct ppa_addr ppa_addr;
159         dna_addr_t dna_ppa_list;
160     };
161
162     struct ppa_addr *ppa_list;
163
164     void *metadata;
165     dna_addr_t dna_metadata;
166
167     uint8_t opcode;
168     uint16_t nr_pages;
169     uint16_t flags;
```



Related work

- FTL na hoście używane przez Baidu, inne implementacje przygotowane przez FusionIO i Violin Memory.
- Przesyłanie wskazówek co do tego jak rozmieścić dane.
- Ciągłe ulepszane kontrolery.

Na podstawie:

<https://www.usenix.org/system/files/conference/fast16/fast16-papers-schroeder.pdf>

<https://arxiv.org/pdf/1706.08642.pdf>

<https://www.usenix.org/system/files/conference/fast17/fast17-bjorling.pdf>

https://pl.wikipedia.org/wiki/Dysk_p%C3%B3%C5%82przewodnikowy