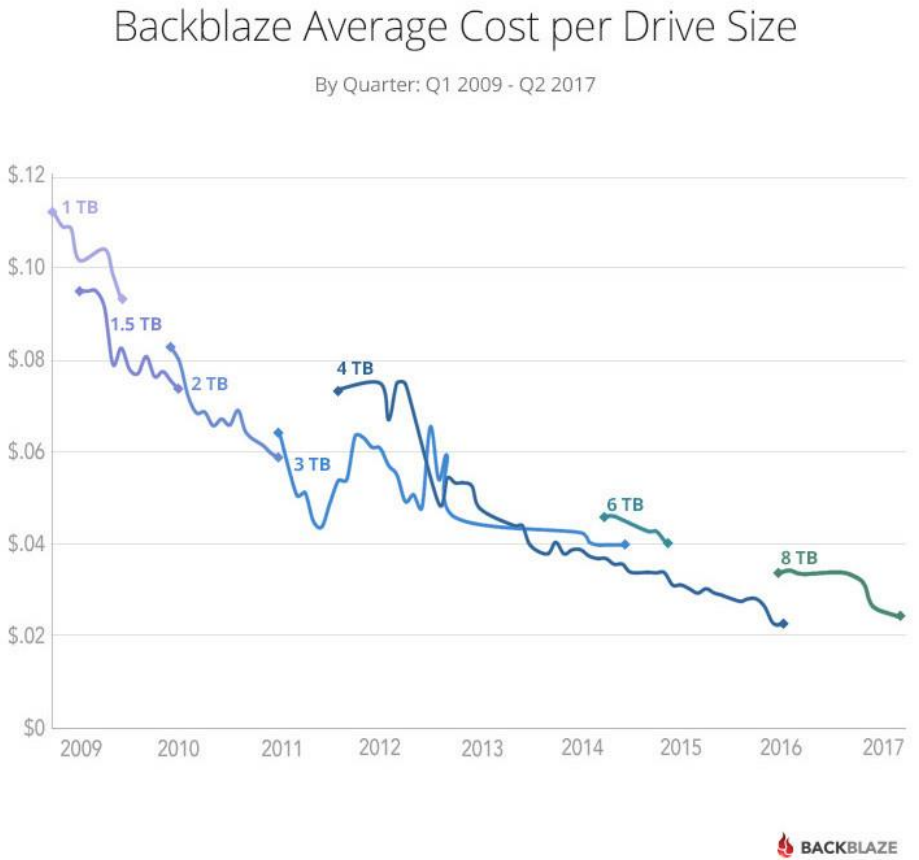
PRINCIPLES OF DATA ORGANISATION

Solid State Drive Introduction

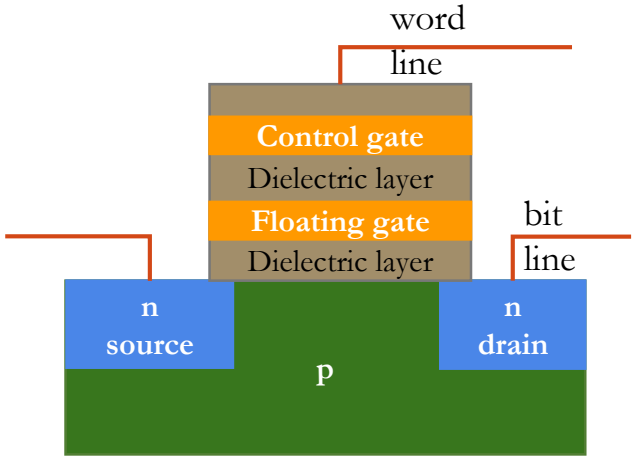
MOTIVATION

- Key, pointer pairs ~ index.
- Solid State Drive (SSD)
- 2018.11.26
  SSDs Are Cheaper Than Ever...



CELL

- No moving mechanical components.
- Flash memory is based on floating gate transistors supporting memory non volatility.
- Floating gate transistors form **floating gates** (cages) capable of holding electrons and the charge they represent.
- If the cell is uncharged it represents a 1, if it is charged it represents a 0. Uncharged gate conducts current.
- A cell always conducts the current when it is energized to a higher than threshold current C_T.
- Multiple cells can then store complex information.



CELL TYPE

- With one charge level, the cell can contain one bit - single-level cell (SLC).
  SLCs are more reliable and less complex but much more expensive, only enterprise solutions contain SLCs.
- With four levels of charge, each cell could contain 2 bits - multi-level cell (MLC).
- Each charge level corresponds to a value
  (e.g., highest charge = 11 … lowest charge = 00).
- Increases storage density and complexity of reading and writing, decreases lifetime.

CELL TYPE FUTURE

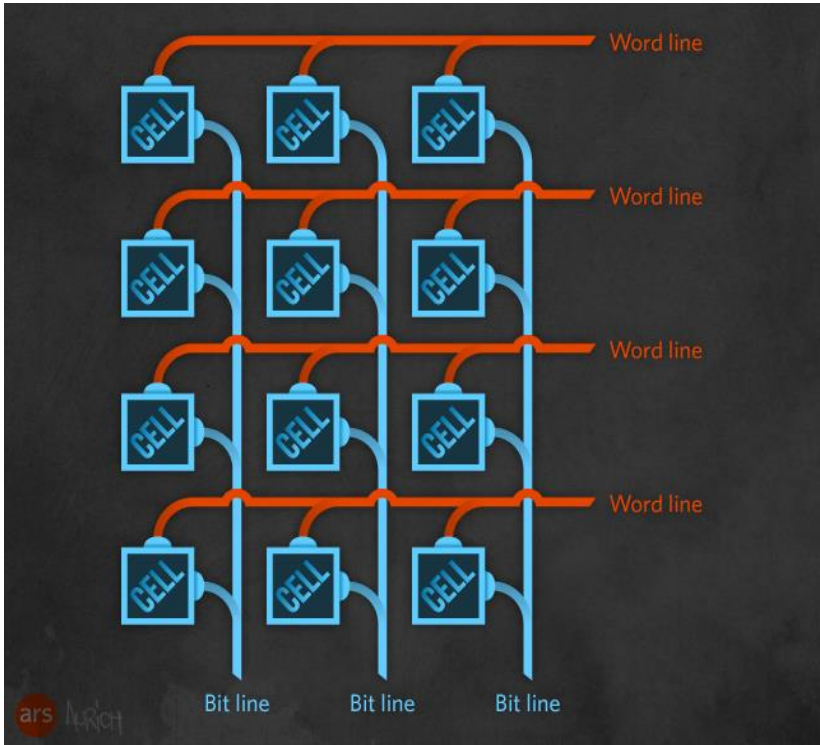
- SLC (Single-Level Cell)
- MLC (Multi-Level Cell)
- TLC (Triple-Level Cell)
- QLC (Quad-Level Cell)
  - 2018 : SSD Micron 5210 Ion
- PLC (Penta-Level Cell)
  - 2019.08.25 : Toshiba Talks 5-Bit-per-Cell Flash
  - 32 distinct voltage level
- ...

BLOCKS AND PAGES

- Cells are stored in a grid called block with rows called pages.
- Pages consists of main memory area and spare area (error correction, management information).
- An SSD consists of multiple blocks.
- Wiring of the cells determines the memory type:
  - NOR memory
  - NAND memory

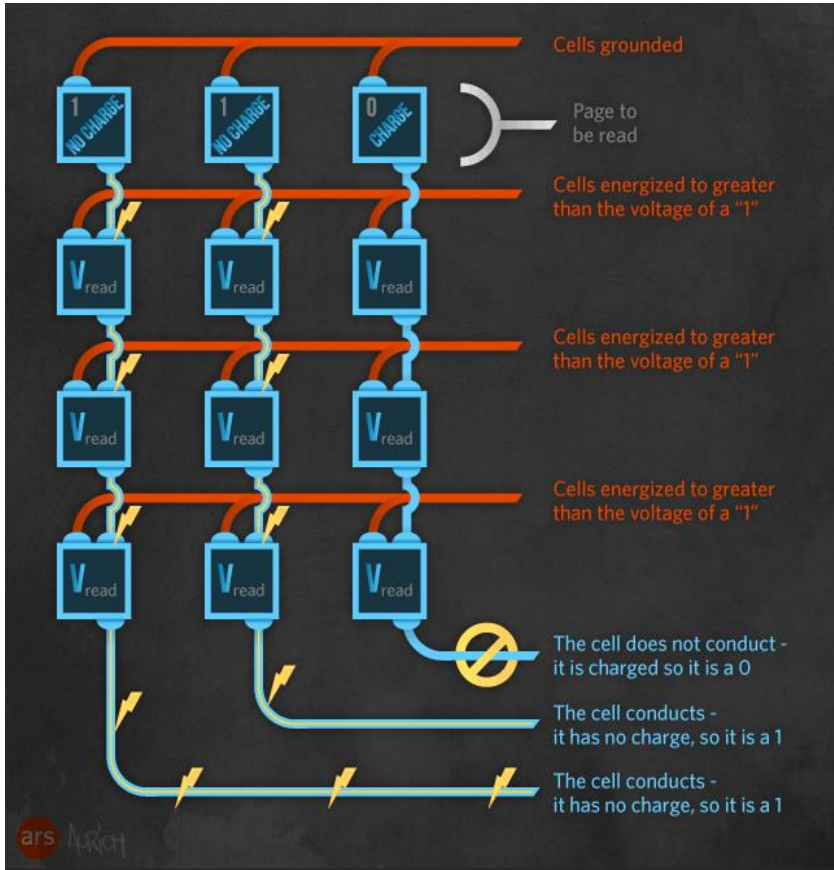
NOR MEMORY

- Each row and column is wired together.
- Application of current to gates and seeing whether the current flows.
- Energize the word line to a low voltage and the bit line will show charge only if the floating gate contains no charge, otherwise if the gate contains a charge the low voltage will not go through.
- NOR chips are complex and take a lot of space, as a result NAND used in SSDs.



NAND MEMORY

- Transistors are connected in series with respect to the bit lines.
- A cell always conducts the current when it is energized to a higher than threshold current C_T.
- All the word lines which are not read are energized to voltage >= C_T.
- The word line to be read is energized with lower current, the current is let in and respective bit lines are checked.
- The bit lines are read simultaneously.



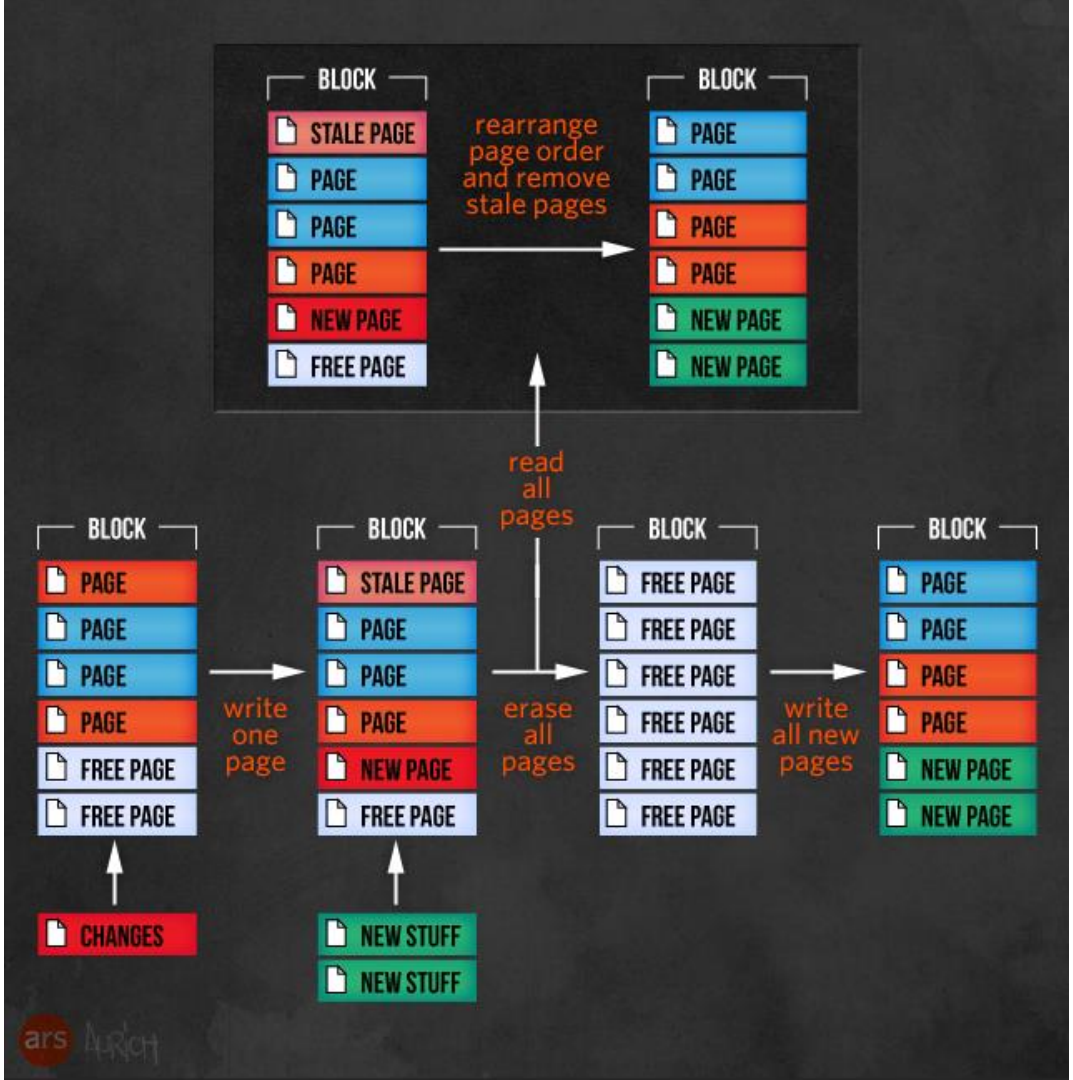
PAGE MODIFICATION

- Smallest addressable unit in SSD is a page.
- The size in modern SSDs uses to be 8,192 bytes.
- Freshly erased page stores all 1s (no charge in the gates) and the cells can be written to on the page level.
- Erasing a page is done by application of high voltages.
- Turning individual cell or page back into 1s is not possible due to the effect on the adjacent cells (high voltage) → erase operation is possible on block level only.
- SSDs therefore do not allow in-place update of data.

PAGE MODIFICATION

- Three types of pages
  - Free page
  - Live/used page
  - Dead/stale page
- Updating a page differs based on whether a free page is available.

- **Free page available**
  - The updated content is written into the new free page.
  - The old pages is marked as dead and can not be used until the whole block is erased
- **No free page available**
  (but some dead/stale page is).
  - The block is read to cache.
  - The block is erased.
  - The modified content is written back.


BLOCK
STALE PAGE
PAGE
PAGE
PAGE
NEW PAGE
FREE PAGE
rearrange page order and remove stale pages
BLOCK
PAGE
PAGE
PAGE
PAGE
NEW PAGE
NEW PAGE
read all pages
BLOCK
PAGE
PAGE
PAGE
PAGE
FREE PAGE
FREE PAGE
CHANGES
write one page
BLOCK
STALE PAGE
PAGE
PAGE
PAGE
NEW PAGE
FREE PAGE
NEW STUFF
NEW STUFF
erase all pages
BLOCK
FREE PAGE
FREE PAGE
FREE PAGE
FREE PAGE
FREE PAGE
FREE PAGE
write all new pages
BLOCK
PAGE
PAGE
PAGE
PAGE
NEW PAGE
NEW PAGE

MEMORY DEGRADATION

- When writing to a page the word line is charged with high voltage and the bit lines which are to be set to 0 are grounded which causes the electrons to migrate into the respective cells.
- Erasing a page is provided by releasing the negative charge from the gate.
- Each cycle causes some residual charge to remain in the cells (damages the dielectric oxide layer) which changes the resistance of the gate. Flipping a gate needs higher current and takes longer.
- Data can be still read but can't be written into the worn-out cells any more.
- In a standard use, the SSD disks should not reach the maximum amount of the program/erase (P/E) cycles sooner than in 5, 10, 24, … years depends also on the level of the NAND memory.

FLASH MEMORY ADDRESSING - SOFTWARE

- Based on log-structured file systems.
- File systems often implement B-trees themselves in order to manage the storage structure.
- JFFS2 (Journalling Flash File System version 2), JFFS3 (wandering tree problem), YAFFS, …

FLASH MEMORY ADDRESSING - HARDWARE

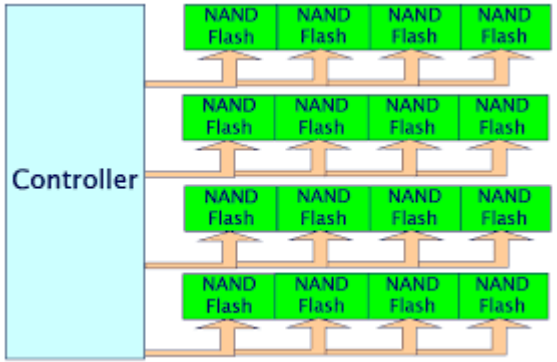
- Interface to the flash chip - Flash Translation Layer (FTL)
- Implemented using a micro-controller within the flash package
- Disk-like interface
  - Simulation of in-place updates by mapping rewrites of a page to an empty page
- Mapping the physical page addresses to logical block addresses and only the logical address is visible outside the package
- Wear levelling by distributing writes uniformly across the media

SSD CONTROLLER

- Processor mediating the communication between SSD memory blocks and the computer.
- Deals with all the logic regarding page management within the NAND chips (including reading, writing, and erasing).
- Tasks:
  - Parallelisation
  - Caching
  - Wear levelling
  - Garbage collection
  - ...

PARALLELIZATION

- SSD has multiple channels and can thus address multiple NAND chips at the same time.
- The controller stripes data in a similar way as the controller in a RAID array does and also provides error correction



CACHING

- When striping is not enough, the controller can use SDRAM to hold data until they can be written to the disk
  - Further decrease in latency
- Requires additional power supply for the volatile SDRAM

WEAR LEVELING

- Keeping track of highly used pages
- Once a time highly and sparsely used pages can be swapped to ensure roughly the same lifespan for all the cells

GARBAGE COLLECTION

- Keeping track of blocks which contain dead/stale pages
- Once a time, blocks with sufficiently enough dead pages are rewritten into newly erased blocks and the old blocks are erased

TRIM

- Existing operating systems do not physically delete files but only remove pointers to them
  - No way for a garbage collector to find out that a given page is dead
- TRIM command
  - These pages are no longer used, when you want, erase them and use
- The TRIM command will only work if the SSD controller, the operating system, and the filesystem are supporting it

CHALLENGES

- Since updating a page needs considerable effort, which is moreover related to the memory degradation, measures need to be taken to decrease the impact of such behavior
  - E.g., decreased number of update operations
- Utilization of controller parallelization capabilities
  - Classical B-tree can not