common

EUROPE LUXEMBOURG

ASSOCIATION OF IBM COMPUTER USERS

Welcome !

EUROPE LUXEMBOURG

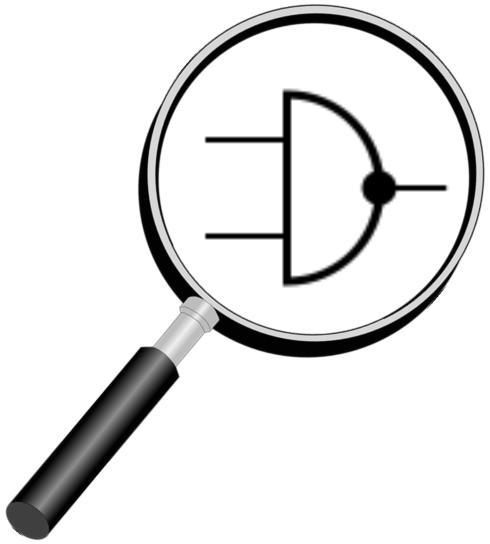
Introduction

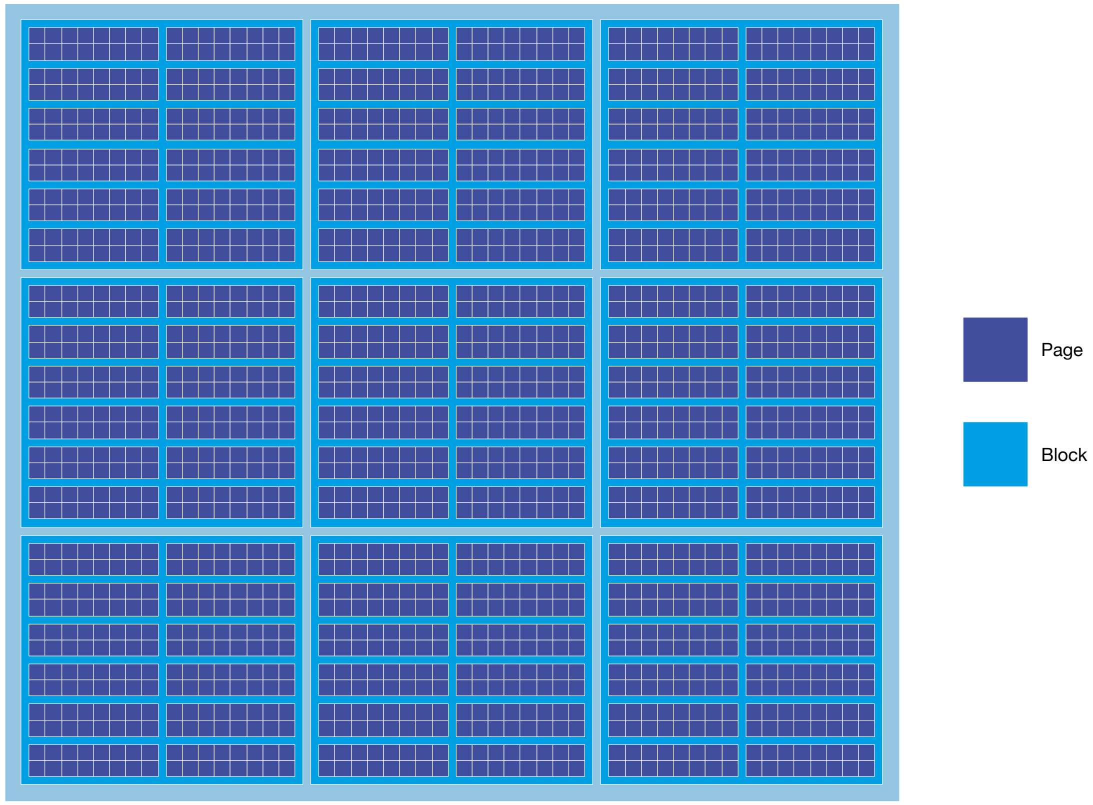
Agenda

- From the Stone Age to the Store Age
- Would you Dare ?
- Back to the Future
- Questions and Answers …

What is Flash Memory ?

- Electronic (solid-state) non-volatile computer storage
  - Can be electronically erased
  - Can be electronically (re)programmed
- Based on cells built on NAND (NOT-AND) gates
- Cells grouped into pages
- Pages grouped into blocks





From the Stone Age to the Store Age

Allowed Operations

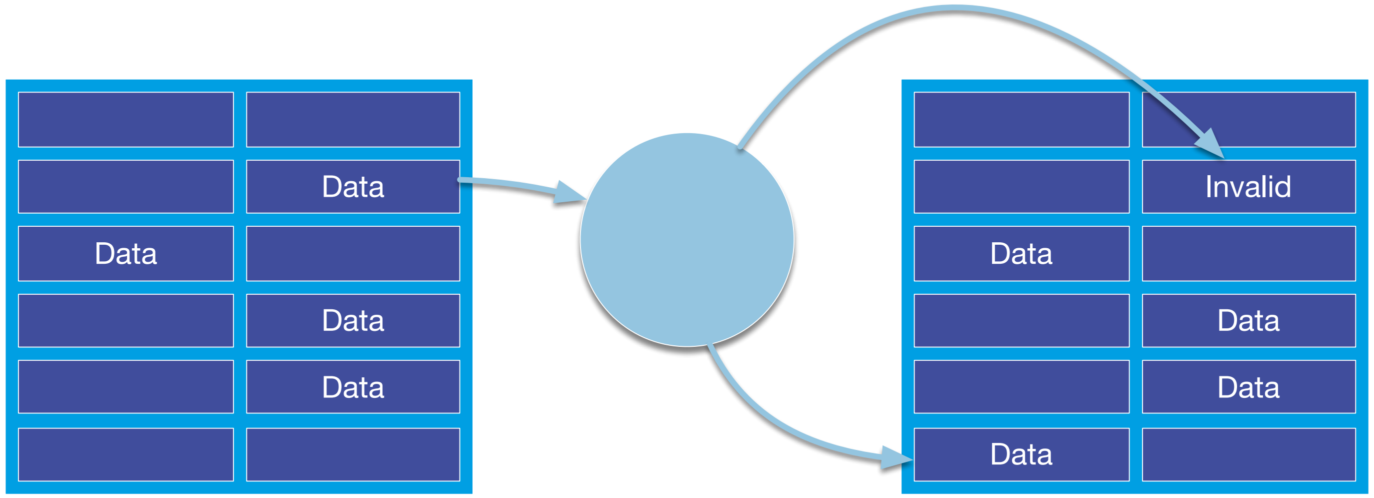
- Read a page
- Program (write) a page
- Erase a block

Disallowed Operations

- “Rewrite” a block or a page
- Erase a page

Techniques

- Over-provisioning
- Bad block detection
- Wear levelling
- Triggering



Flash Memory : Terminology Used

- SSD = Solid State Drive
  - Device that uses solid-state storage to store data persistently
  - Please don’t call it anymore a “disk” !
- Program/Erase (P/E) cycle
  - When data is written to a cell, erased, and re-written
- SLC = Single Level Cell
  - Single bit value per cell (2 values)
  - Longest lifespan, most expensive
  - Supports up to 100.000 P/E cycles
- MLC = Multi Level Cell
  - Two bits of data per cell (4 values)
  - Supports up to 10.000 P/E cycles
- eMLC = Enterprise Multi Level Cell
  - Enhanced controller logic, error recovery, construction density
  - Supports up to 30.000 P/E cycles

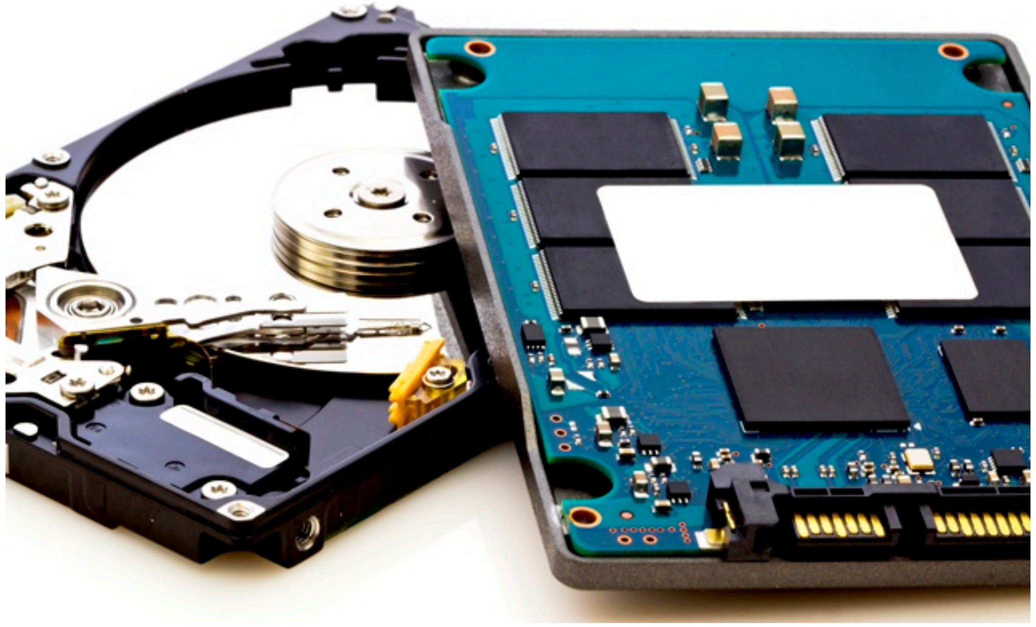
Flash Memory : Terminology Used

- TLC = Triple Level Cell
  - Three bits of data per cell (8 values)
  - Higher requirements for error correction and wear levelling
  - Supports up to 5.000 P/E cycles
- 3DTLC
  - TLC organised in spatial layers (X, Y and Z axis)
  - From 32 to 48 layers …
  - From 48 to 64 layers …
- QLC = Quad Level Cell
  - Four bits of data per cell (16 values)
  - Currently in early deployment …
- PLC = Penta Level Cell
  - Five bits of data per cell (32 values)
  - Currently in development …

From the Stone Age to the Store Age

Performance

- Hard Disk Drive
  - Enterprise 10k and 15k RPM
  - Performance stays around 200 I/O per second
  - Power consumption
    - ★ 2.5” : 0.7 to 3.0 Watts
    - ★ 3.5” : 6.5 to 9.0 Watts
- Flash Memory
  - No mechanical moves
  - No rotational delay
  - Lower latency
  - Higher IO/s
  - Performance > 1.000.000 I/O per second
  - Power consumption
    - ★ 0.6 to 1.8 Watts



About “Read-Intensive” (aka. “Mainstream”) Flash Memory

- What is it ?
    - Lower endurance solid-state manufacturing
        - ★ Use of MLC/TLC instead of SLC/eMLC
        - ★ Lower over-provisioning
    - Lower cost
    - Lower write performance
- DWPD = Drive Write Per Day
    - Highest endurance : support up to 30 DWPD
    - Enterprise @ IBM : support up to 10 DWPD
    - Enterprise @ OEM : support up to 3 DWPD
    - Mainstream : support 1 DWPD
    - Laptop SSD : supports up to 0.3 DWPD
    - USB Stick : support up to 0.1 DWPD

About “Read-Intensive” (aka. “Mainstream”) Flash Memory

- Recommandations
  - Do not mix read-intensive drives with mainstream drives in disk arrays
  - Do not use read-intensive drives for easy-tiering
  - Monitoring end of life for read intensive drives
    - ★ Predictive Failure Analysis (PFA)
    - ★ Using the fuel gauge command
  - Plan for RAID-6 or DRAID-6 !
  - Plan for over-provisioning
  - Plan for spares !

From the Stone Age to the Store Age

IBM FlashCore Technology ?

- The DNA of the IBM FlashSystem Family
- Able to monitor individual flash cells
  - Extremely low latency
  - Predictive Techniques
- Unprecedented capacity
  - High performance compress/decompress algorithms
  - Compression came from IBM Mainframe
  - Minimize data written to flash
  - Data reduction is transparent
- Modules (raw) capacities
  - 4.8 TB
  - 9.8 TB
  - 19.2 TB
  - 38.4 TB
- Extreme endurance
  - 10 DWPD !
  - Chip-level RAID on modules (VSR)
- Complexity of firmware

From the Stone Age to the Store Age

IBM SCM (Storage Class Memory) ?

- The cache/memory/storage hierarchy is rapidly becoming the bottleneck for large systems
- Speeds Paradigm
  - CPU : 1 ns
  - CPU Cache : < 5 ns
  - RAM : 60 ns
  - FCM : < 100 µs
  - SSD : < 1 ms
  - HDD : < 5 ms
  - Tape : 40 s
- Human Perspective
  - CPU : second
  - CPU Cache : second
  - RAM : minute
  - Storage : month
  - Tape : millenium
- Goal of SCM ?
  - Fulfil the gap between memory and storage

IBM SCM (Storage Class Memory) ?

- A “new” device …
  - …

From the Stone Age to the Store Age

NVMe or NVM Express

- Non-Volatile Memory (Host Controller Interface) Specification
  - Protocol created to accelerate the transfer between hosts and storage
  - Over high-speed PCIe Bus
- Legacy design
  - PATA
  - SATA
  - SCSI
  - SAS
- New design
  - More efficient interface
  - Lower latency
  - More scalable
- NVMe-oF
  - New kind of transport to allow NVMe from host to storage

Limitations of Flash

- Asymmetric performance
  - Program/erase cycle
- Endurance
  - Single level cell (SLC) :10^5 writes/cell
  - Multi level cell (MLC) : 10^4 writes/cell
  - Triple level cell (TLC) : ~300 writes/cell
  - …

Flash Storage versus SSD ?

- Solid State Drive
  - Flash memory
  - Accessed through SAS controller/device interface chain
    - ★ SAS-2 : 6 Gbps, ~ 750 MB/s
    - ★ SAS-3 : 12 Gbps, ~ 1500 MB/s
    - ★ SAS-4 (future) : 22.5 Gbps, ~ 2800 MB/s
- Flash Storage
  - Flash memory
  - Accessed directly through PCIe bus !
    - ★ PCIe 2 : 500 MB/s (x1) - 8000 MB/s (x16)
    - ★ PCIe 3 : 985 MB/s (x1) - 15750 MB/s (x16)
    - ★ PCIe 4 : 1969 MB/s (x1) - 31510 MB/s (x16)
  - Micro latency : from milliseconds to microseconds
- One second is …
  - One thousand milliseconds (ms)
  - One million microseconds (µs)
- Short comparison
  - HDD : 5 ms
  - SSD : 1 ms (÷5)
  - Flash : 100 µs (0,1 ms … ÷10 … ÷50) !

From the Stone Age to the Store Age

Target Applications

- Databases
- Virtual Desktop Infrastructure
- Latency Sensitive Apps

Technologies …

	SAS HDD	SAS SSD	NVMe SSD	NVMe FCM	NVME SCM
Type of Media	Rotating Media	3D NAND	3D NAND	3D NAND	
Protocol	SCSI	SCSI	NMVe	NMVe	NMVe
Physical size	2.5” 3.5”	2.5“	2.5“	2.5“	2.5“
Capacities	2 TB - 20 TB	1.9 TB - 30 TB	800 GB - 15.4 TB	4.8 TB - 38.4 TB	375 GB - 1.6 TB
Differentiator				Compression Encryption	Very low latency
Speed	3ms - 5ms	< 1ms	150µs - 250µs	70µs - 100µs	15µs

Would you Dare ?

Performance of “All-Flash” configuration would allow

- Distributed RAID-6
- Encryption
- Compression/deduplication
- Snapshotting
- …

Would you Dare ?

Distributed RAID (DRAID)

- Distributed ?
  - Protection capacity distributed over all drives
  - Spare capacity distributed over all drives
  - No dedicated spare drive : everyone works !
- Advantages ?
  - Better performance
  - Faster rebuilt time
- DRAID-5
  - Stripe data over all the members
  - One parity strip for every data stripe
  - Tolerate the failure of one member drive
- DRAID-6
  - Stripe data over all the members
  - Two parity strips for every data stripe
  - Tolerate the failure of two member drives
- Recommendation
  - Use (D)RAID-6 when unit capacity is over 1 TB !
  - Drive rebuilt is sufficiently long to encounter a second failure …

Would you Dare ?

RAID 5 versus RAID 6

- RAID 5 provides good protection
  - Drive capacities are an issue
  - A second failure is disastrous
- RAID 6 provides better protection
  - Two simultaneous failures ?
- RAID 6 is the better choice
  - To be strongly recommended for units above 1 TB !
  - Must be mandatory for full flash configurations
- DRAID is better than RAID
  - Definitely !
  - Beware of constraints …

Would you Dare ?

Encryption Support

- Encryption-capable
  - Optional (chargeable) capability of a device to encrypt data by using a secret key
- Encryption-disabled
  - No secret key is configured
  - Note that FlashCore devices always encrypt data with an IBM well-know key
- Encryption-enabled
  - A secret key is configured
  - The device encrypts user and metadata with that key
- Access-control-enabled
  - An access key must be provided to access an encrypted entity
- Protection-enabled
  - Encryption-enabled
  - Access-control-enabled
- Protection Enablement Process (PEP)
  - Performed during storage device initialisation process
- Application transparent encryption

Would you Dare ?

(Real-time) Compression

- Data storage reduction technology
  - Inline, lossless data compression
- When/where used ?
  - Active primary data and/or replicated/mirrored data
  - General-purpose, database, virtualized infrastructures volumes
- DOs and DONTs
  - Best candidates are data type not compressed by nature
    - ★ Database and/or character data
    - ★ E-mail systems
    - ★ Vector data (CAD/CAM)
  - Worse candidates are
    - ★ Compressed audio/video (JPEG, MPEG, …)
    - ★ Compressed user productivity formats (DOCX, PPTX, XLSX, …)
    - ★ Other compressed formats (TAR, ZIP, …)
    - ★ Encrypted data

- Used on following IBM products
  - Spectrum Virtualize
    - ★ SAN Volume Controller
    - ★ Storwize V7000 Gen2 and V5030 Gen2 (without compression accelerators)
- Recommendation

common
EUROPE LUXEMBOURG

Would you Dare ?

Deduplication

- Data storage reduction technology
- When/where used ?
  - Effective when highly redundant data sets can be found
    - ★ Backup
    - ★ Virtualization
  - Relies on a (highly solicited) database to store pointers
    - ★ Performance access relies on the health of that database
    - ★ Data integrity relies on the health of that database
    - ★ Why not DRAID-6 on Flash ?
- Recommendation
  - Use Data Reduction Estimator Tool

Back to the Future …

IBM FlashSystem 5010

- Entry solution, built-in with Spectrum Virtualize
- Capacities
  - Cache : 16, 32 or 64 GB
  - 392 Drives
- External Connectivity
  - FC, iSCSI, iWARP, RoCE, SAS
- Internal Connectivity
  - SAS
- Maximum 10 expansions
- No Cluster
- Can be hybrid

Back to the Future …

IBM FlashSystem 5030

- Entry solution, built-in with Spectrum Virtualize
- Capacities
  - Cache : 32 or 64 GB
  - 504 Drives
- External Connectivity
  - FC, iSCSI, iWARP, RoCE, SAS
- Internal Connectivity
  - SAS, NVMe
- Maximum 20 expansions
- 2-Ways Cluster
- Can be hybrid

Back to the Future …

IBM FlashSystem 5100

- Entry solution, built-in with Spectrum Virtualize
- 24x NVMe slots in the control unit
- Capacities
  - CPU : 2x 8-Cores
  - Cache : 64 to 576 GB
  - 504 Drives
  - Maximum I/Os : 900000
  - Maximum Throughput : 15 GB/s
- External Connectivity
  - FC, FC-NVMe, iSCSI, iWARP, RoCE, SAS
- Internal Connectivity
  - SAS, NVMe
- Maximum 20 expansions
- 2-Ways Cluster
- Can be hybrid

Back to the Future …

IBM FlashSystem 7200

- Midrange solution, built-in with Spectrum Virtualize
- 24x NVMe slots in the control unit
- Capacities
  - CPU : 4x 8-Cores
  - Cache : 256 to 1536 GB
  - 504 Drives
  - Latency : < 70 µs
  - Maximum I/Os : 2300000
  - Maximum Throughput : 35 GB/s
- External Connectivity
  - FC, FC-NVMe, iSCSI, iWARP, RoCE, SAS
- Internal Connectivity
  - SAS, NVMe
- Maximum 20 expansions
- 4-Ways Cluster
- Can be hybrid

Back to the Future …

IBM FlashSystem 9200/9200R

- High-end solution, built-in with Spectrum Virtualize
- 24x NVMe slots in the control unit
- Capacities
  - CPU : 4x 16-Cores
  - Cache : 256 to 1536 GB
  - 504 Drives
  - Latency : < 70 us
  - Maximum I/Os : 4500000
  - Maximum Throughput : 45 GB/s
- External Connectivity
  - FC, FC-NVMe, iSCSI, iWARP, RoCE, SAS
- Internal Connectivity
  - SAS, NVMe
- Maximum 20 expansions
- 4-Ways Cluster
- Can be hybrid

Questions & Answers

EUROPE LUXEMBOURG

Thank You !

common

EUROPE LUXEMBOURG

common

EUROPE LUXEMBOURG