ISOTDAQ

International School of Trigger and Data Acquisition

Storage systems for DAQ

~~Adam Abed Abud (CERN)~~ Enrico Gamberini (CERN)

ISOTDAQ 202~~0~~2

13 - 23 June 2022 (Catania, Italy)

Storage Examples in Bytes

Google global storage
(10-15 EB)

CERN
(110 PB/year)

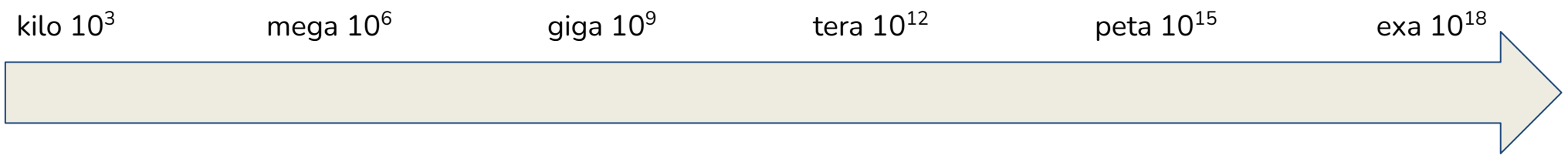
DUNE to storage
(250 MB/s)

DUNE pre-trigger
(1.5 TB/s)

DUNE to storage
(7.5 PB/year)

4K video stream
(4 MB/s)

1 hour of video
(1-10 GB)



Storage Examples in Bytes

Google global storage
(10-15 EB)

CERN
(110 PB/year)

YouTube to storage
(3 GB/s)

YouTube to storage
(90 PB/year)

ATLAS to storage
(1-5 GB/s)

ATLAS pre-trigger
(60 TB/s)

ATLAS to storage
(20 PB/year)

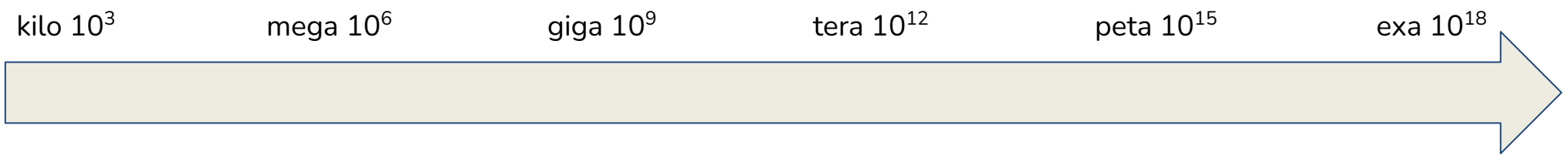
DUNE to storage
(250 MB/s)

DUNE pre-trigger
(1.5 TB/s)

DUNE to storage
(7.5 PB/year)

4K video stream
(4 MB/s)

1 hour of video
(1-10 GB)

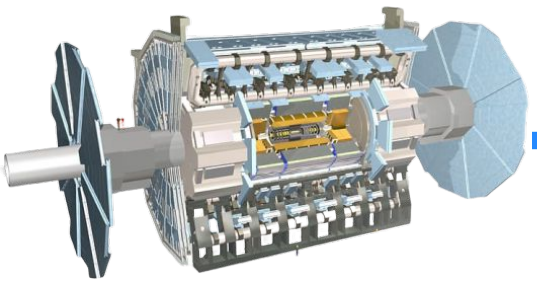


Outline

- Why are storage systems relevant for DAQ ?
- Storage concepts
- Technology overview
    - HDD, SSD, NVM and DRAM
- Performance benchmarking
    - DD and FIO
- Storage challenges for the future
- R&D for DUNE: Supernova burst trigger
- Conclusion

Why are storage systems relevant for DAQ ?

TDAQ pipeline



Detector

40 MHz

1 MB/evt

100 kHz

L1 Trigger

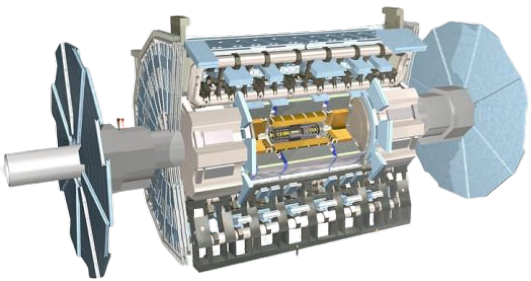
1 kHz

High-Level Trigger

Physics analysis

Why are storage systems relevant for DAQ ?

TDAQ pipeline



Detector

40 MHz

1 MB/evt

100 kHz

1 kHz

L1 Trigger

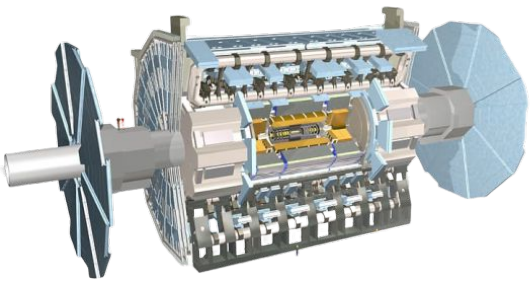
High-Level Trigger

Physics analysis

- Not all the data can be stored:
  - Lack of storage resources
  - Not enough (offline) processing power

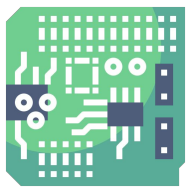
Why are storage systems relevant for DAQ ?

TDAQ pipeline



40 MHz

1 MB/evt



100 kHz

1 kHz

Detector

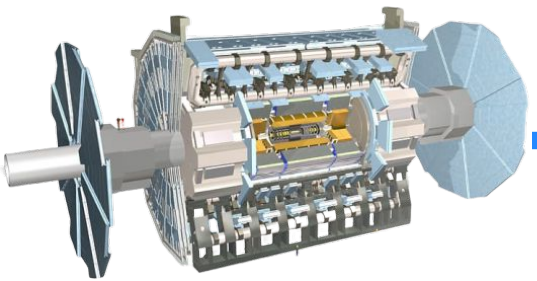
L1 Trigger

High-Level Trigger

Physics analysis

Why are storage systems relevant for DAQ ?

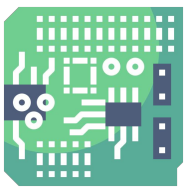
TDAQ pipeline



Detector

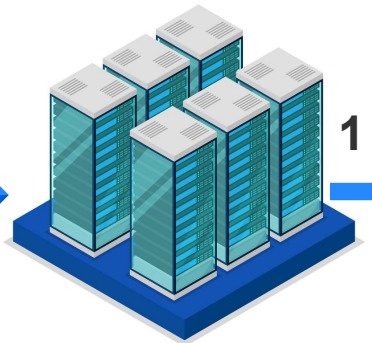
40 MHz

1 MB/evt



L1 Trigger

100 kHz



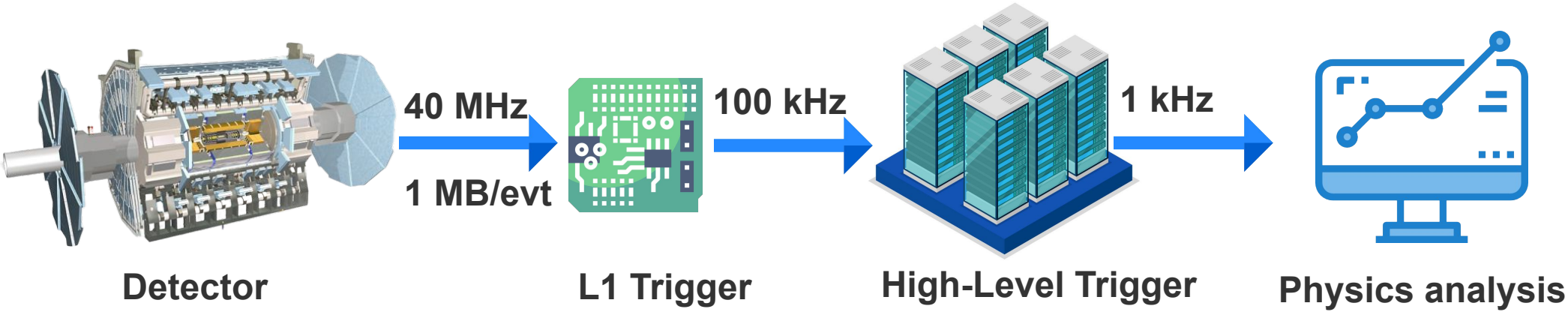
High-Level Trigger

1 kHz

Physics analysis

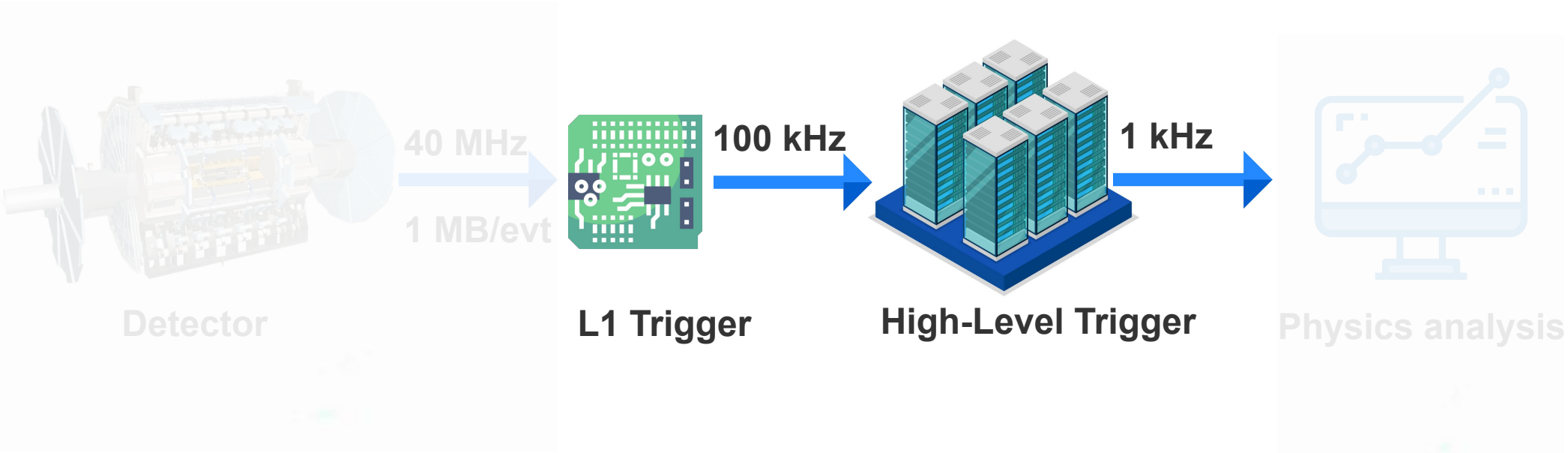
Why are storage systems relevant for DAQ ?

TDAQ pipeline and physics analysis



Why are storage systems relevant for DAQ ?

TDAQ pipeline - Online data taking (“DAQ”)



“Safely store data from point A to point B”

DAQ takeaway

Online vs Offline

- Storage systems ensure that data is stored and physics results can be produced!
  - Data stored → physics results
- DAQ requirements are different from offline analysis:
  - Storage used to buffer data:
    Absorbs rate fluctuations from the rest of the system
  - Continuous stream of data flow **in and out**
    the storage system
  - **Throughput** and **latency constraints**
  - Technology choice affected by **total expected data**

DAQ takeaway

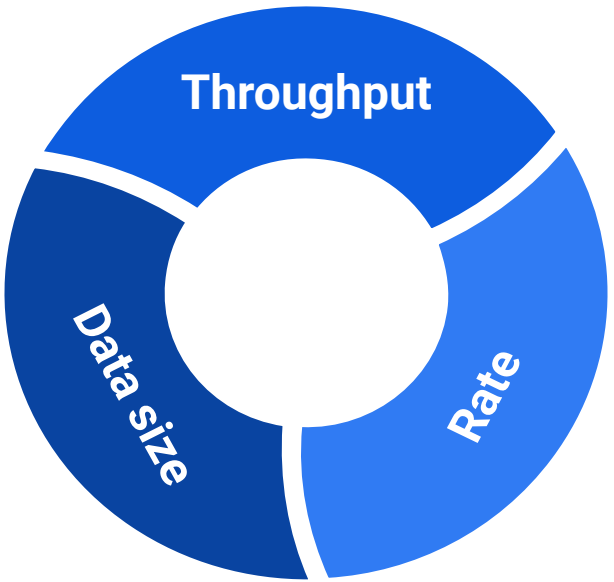
Online vs Offline

- Storage systems ensure that data is stored and physics results can be produced!
  - Data stored → physics results
- DAQ requirements are different from offline analysis:
  - Storage used to buffer data:
    Absorbs rate fluctuations from the rest of the system
  - Continuous stream of data flow **in and out**
    the storage system
  - **Throughput** and **latency constraints**
  - Technology choice affected by **total expected data**

DAQ takeaway

Online vs Offline

- Storage systems ensure that data is stored and physics results can be produced!
  - Data stored → physics results
- DAQ requirements are different from offline analysis:
  - Storage used to buffer data:
    Absorbs rate fluctuations from the rest of the system
  - Continuous stream of data flow **in and out**
    the storage system
  - **Throughput** and **latency constraints**
  - Technology choice affected by **total expected data**

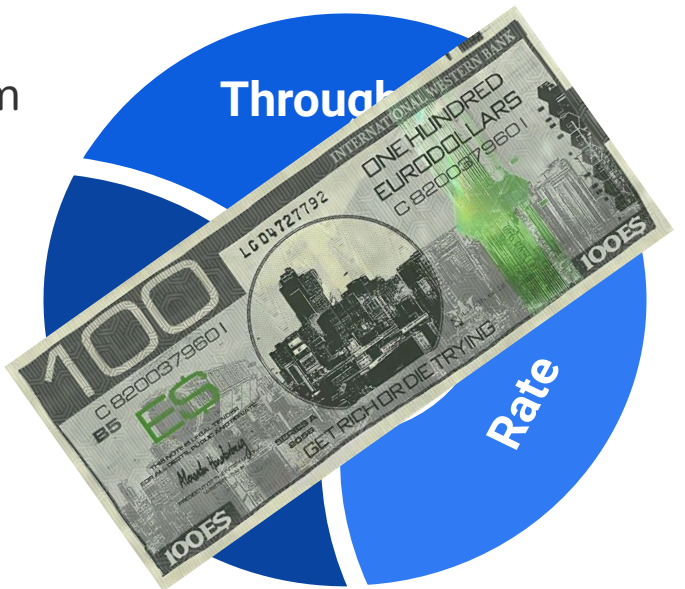


DAQ takeaway

Online vs Offline

- Storage systems ensure that data is stored and physics results can be produced!
  - Data stored → physics results
- DAQ requirements are different from offline analysis:
  - Storage used to buffer data:
    Absorbs rate fluctuations from the rest of the system
  - Continuous stream of data flow **in and out**
    the storage system
  - **Throughput** and **latency constraints**
  - Technology choice affected by **total expected data**

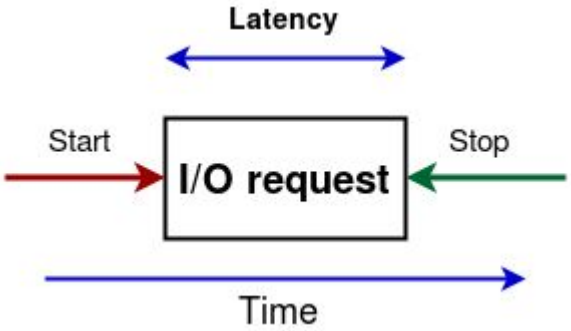
and cost!



Storage concepts and Technology overview

Storage concepts

Some definitions

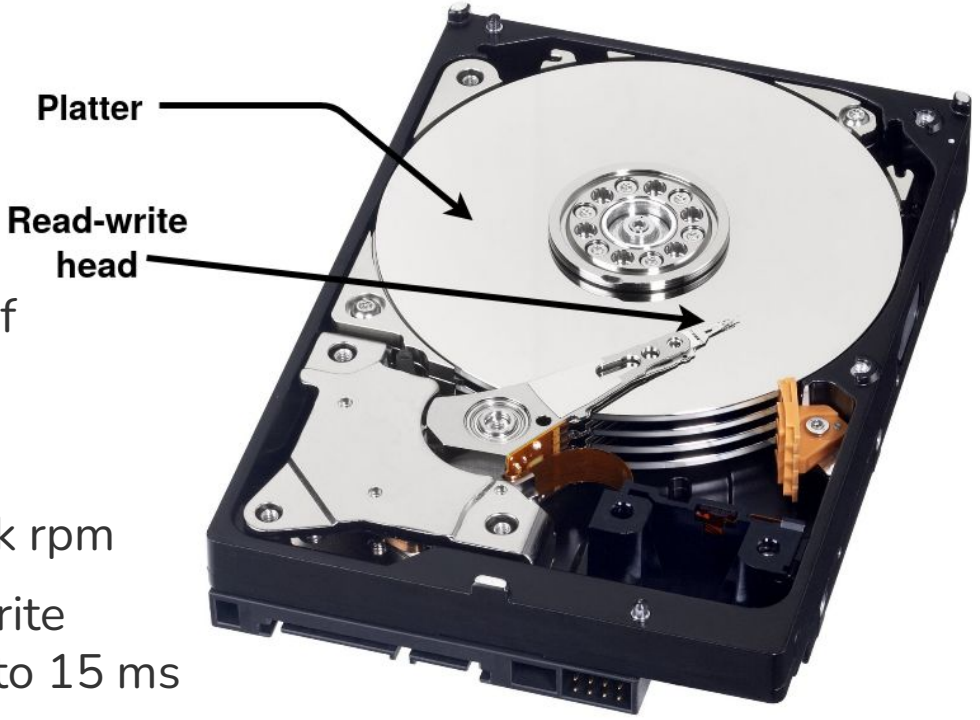


- **I/O:** input/output operation
- **Access pattern:** sequential/random read or write
- **Latency**: time taken to respond to an I/O. Usually measured in ms or in μs
- **Rate:** number of I/O per second to a storage location (**IOPS**)
- **Blocksize:** size in bytes of an I/O request
- **Bandwidth**: product of I/O block size and IOPS

Bandwidth = [I/O block size] x [IOPS]

Hard drives (HDD)

Quick introduction

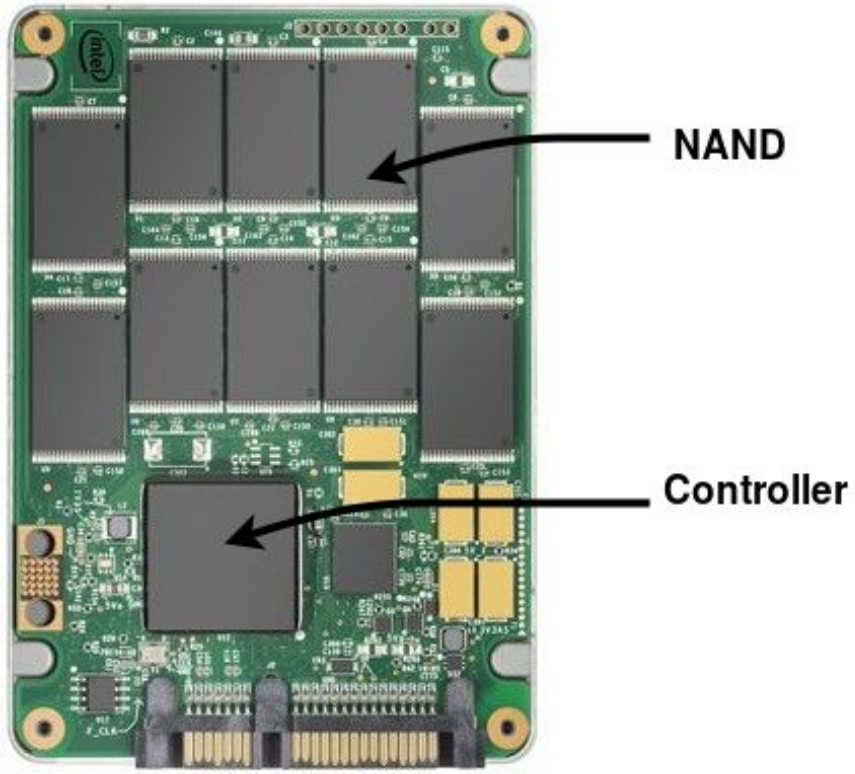


- Electromechanical device
- Circular rotating platter divided into millions of magnetic components where data is stored
- Typical rotational speed of HDDs:
  - 5400 rpm, **7200 rpm**, 10k rpm and 15k rpm
- **Seek time:** time required to adjust the read-write head on the platter. Typical values: from 3 ms to 15 ms
- **Rotational latency:** time needed by the platter to rotate and position the data under the read-write head

$$IOPS = \frac{1}{\text{Avg. seek} + \text{Avg. latency}}$$

Solid state drives (SSD)

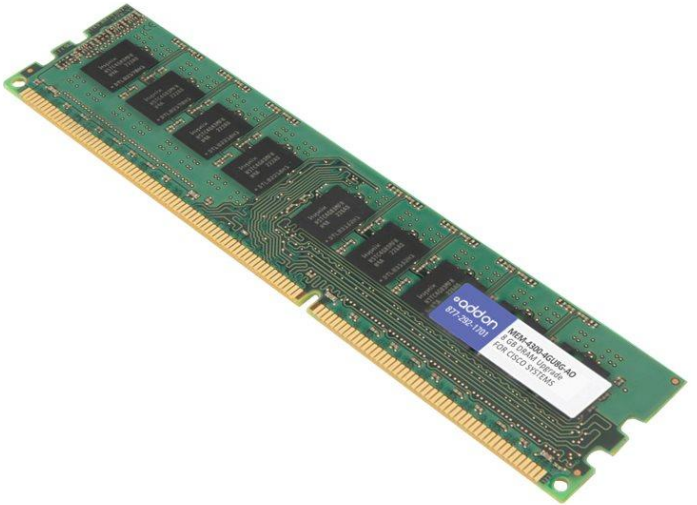
Quick introduction



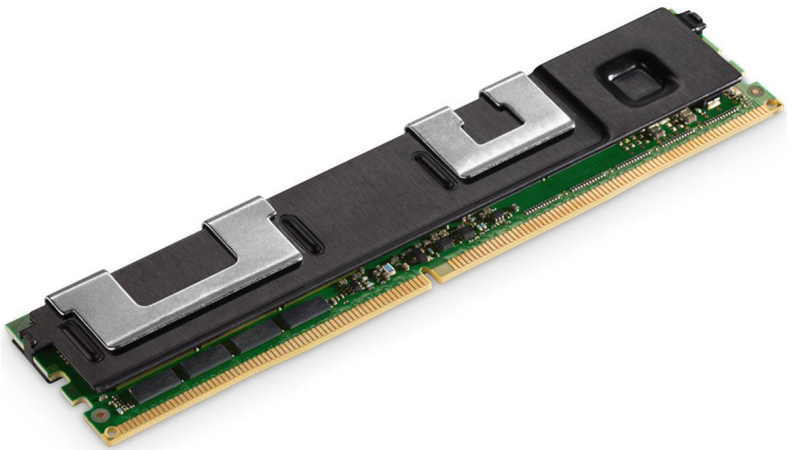
- **Architecture**:
  - NAND flash chipset: store data
  - Controller: caching, load balancing and error handling
- Capacity limited to number of NAND chipsets a manufacturer is able to insert into a device
- (Typically) better performance compared to HDDs
  - There is no mechanical component
  - Reduced latency and seek time
- Optimized controller and communication technology for higher bandwidth devices
  - NVM Express (NVMe) SSD

DRAM and Non-Volatile Memory

Quick introduction

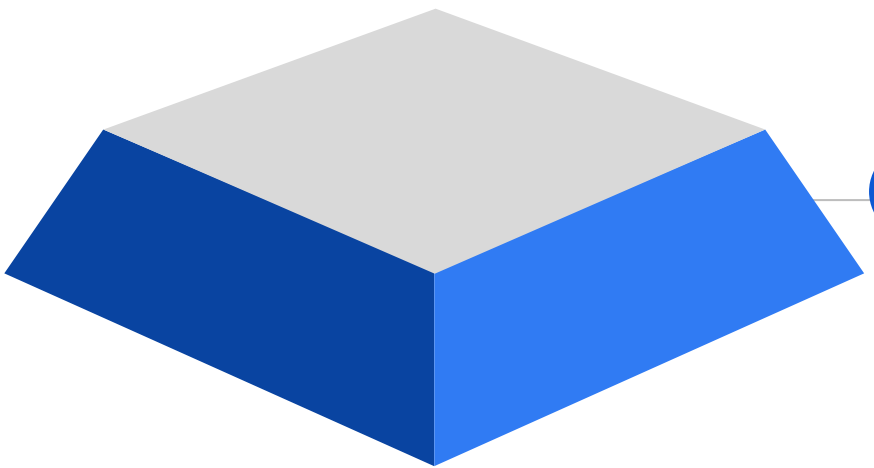


- **DRAM**
  - Semiconductor memory technology
  - Data is not persisted, only temporary storage cells (capacitors and transistors)
  - Low latency (0.1 µs)
- **Non-volatile memory (NVM)**
  - Hold data even if device is turned off
  - Higher storage capacity than DRAM
  - Latency (1 µs)
  - 3D XPoint technology (Intel and Micron, 2015)



Latency and Bandwidth

Technology overview



5

HDD

100 MB/s

10 ms

Latency

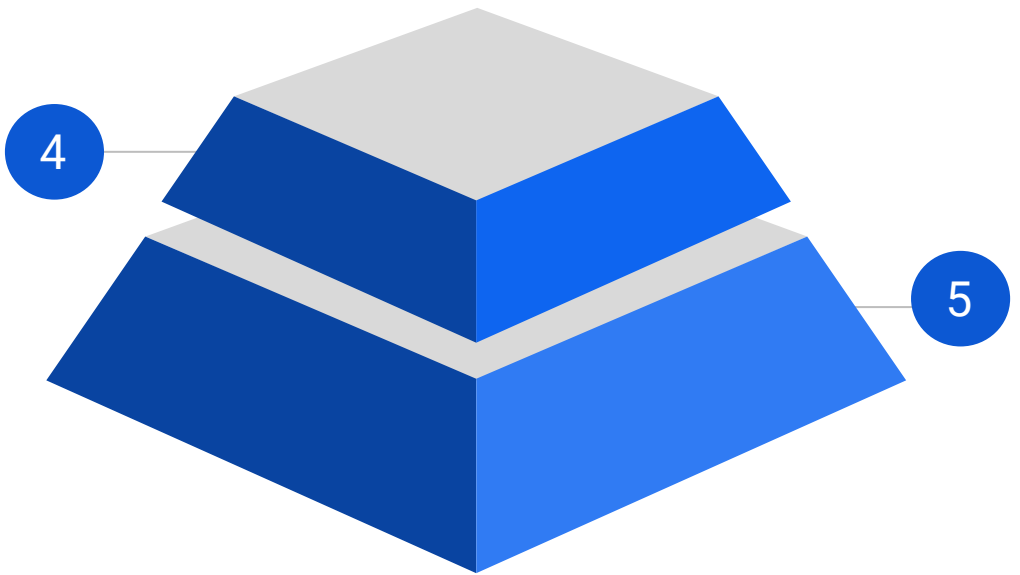
Latency and Bandwidth

Technology overview

SSD

500 MB/s

100 µs



HDD

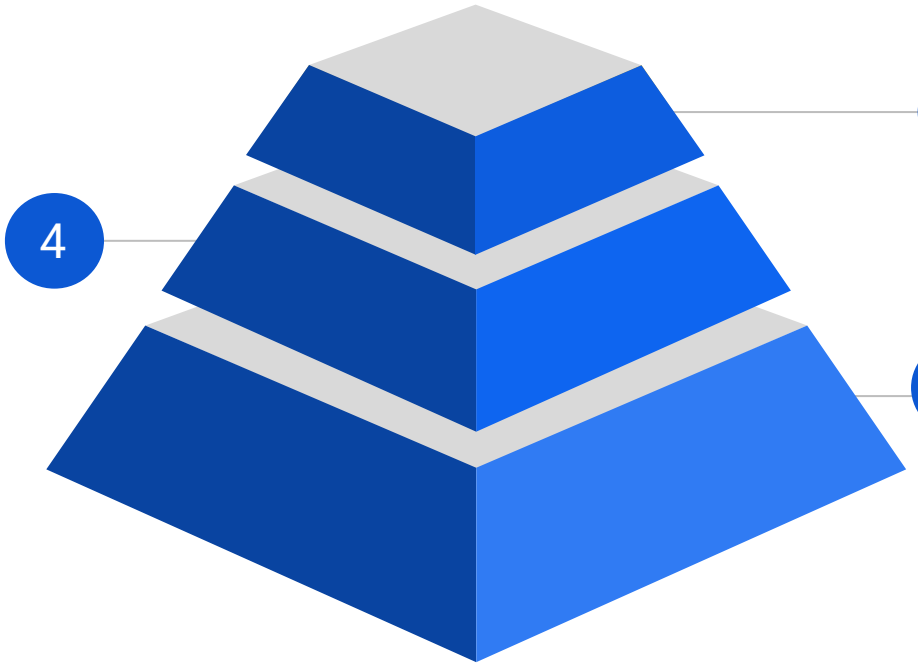
100 MB/s

10 ms

Latency

Latency and Bandwidth

Technology overview



NVMe SSD
2 GB/s
30 µs

SSD
500 MB/s
100 µs

HDD
100 MB/s
10 ms

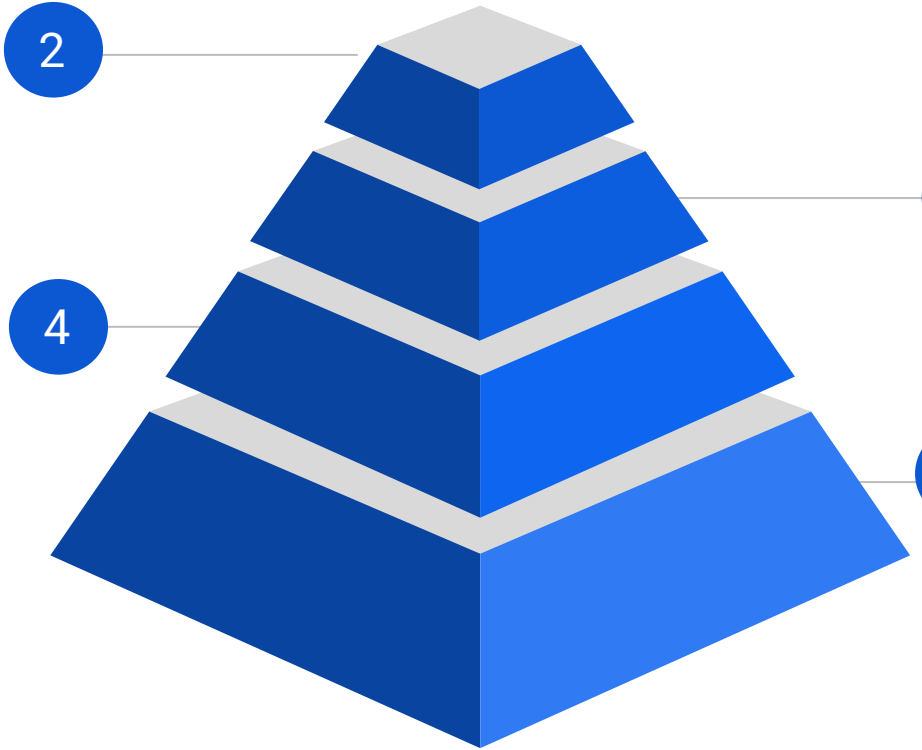
Latency

Latency and Bandwidth

Technology overview

NVM
10 GB/s
1 µs

SSD
500 MB/s
100 µs



NVMe SSD
2 GB/s
30 µs

HDD
100 MB/s
10 ms

Latency

Latency and Bandwidth

Technology overview

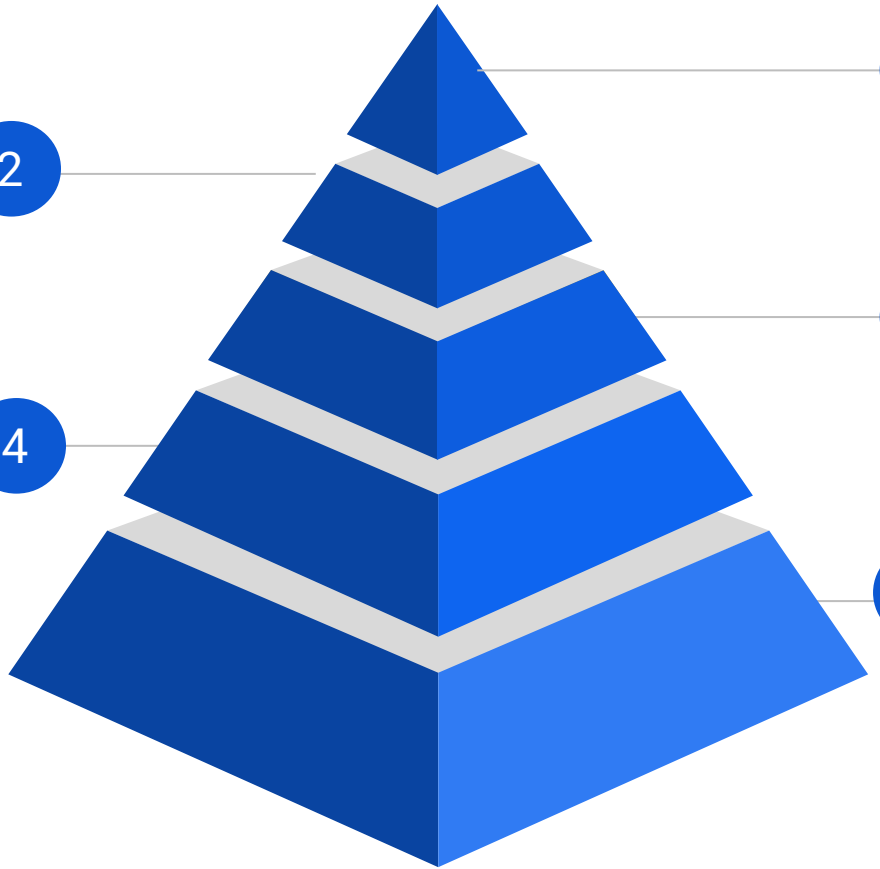
DRAM
30 GB/s
0.1 µs

NVM
10 GB/s
1 µs

NVMe SSD
2 GB/s
30 µs

SSD
500 MB/s
100 µs

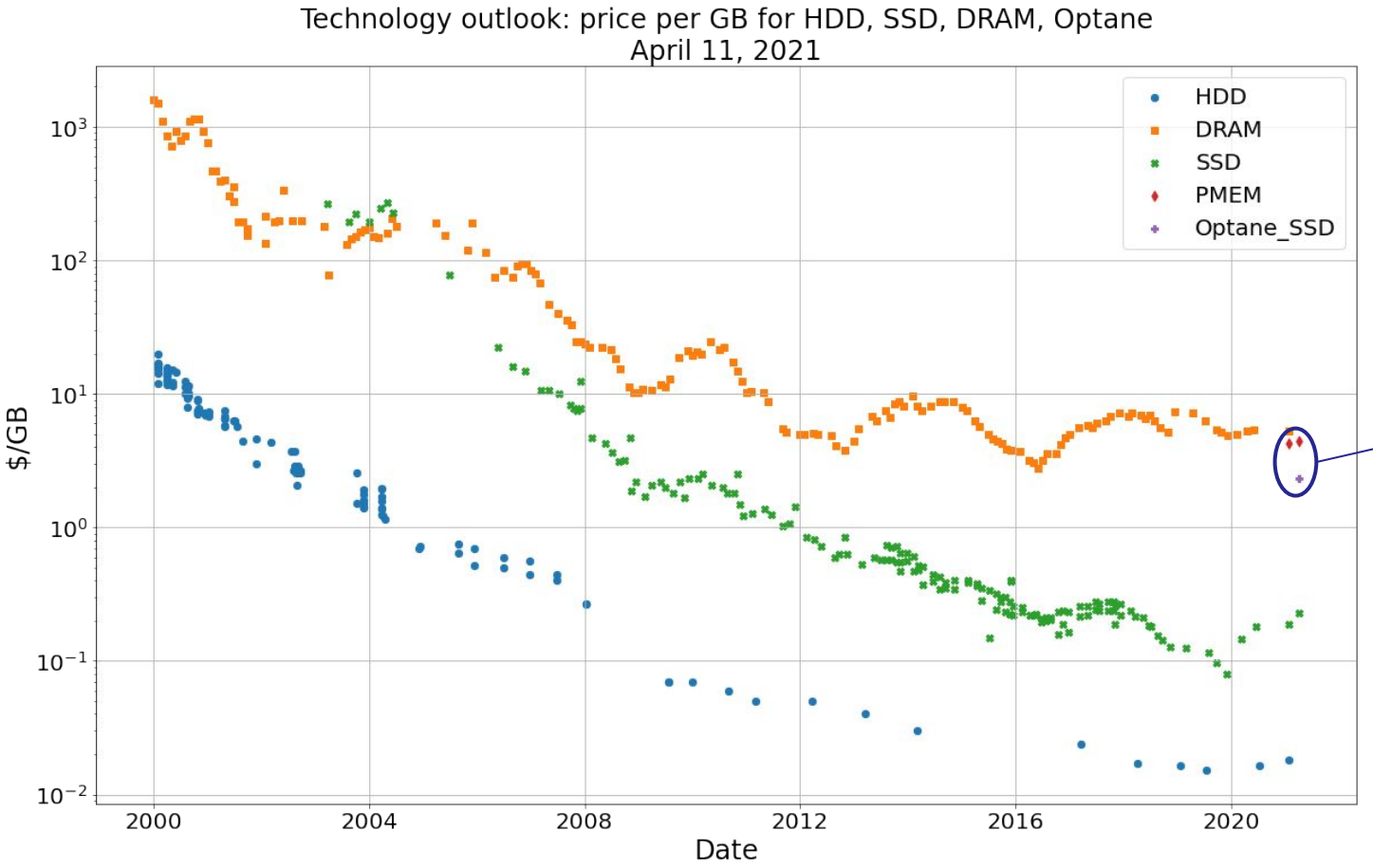
HDD
100 MB/s
10 ms



Latency

Market trend for storage technologies

Price per GB for HDD, SSD, Flash and RAM



3D XPoint (NVM)

Data collected by John C. McCallum.
Since 2018 data collected by Adam Abed Abud

Storage benchmarking

DD

- Linux tool to copy data at the block level
- Usage:
  - `dd if=/path/to/input/file of=/path/to/output/file bs=block_size count=amount_blocks`
- Avoid operating system cache by adding **oflag=direct** option

```
[student@storage_lecture]$ dd if=/dev/zero of=deleteme bs=1M count=1000
1000+0 records in
1000+0 records out
1048576000 bytes (1.0 GB, 1000 MiB) copied, 3.67626 s,  285 MB/s
```

Storage benchmarking

Flexible I/O (FIO)

- Advanced tool for characterizing I/O devices
- Usage:
  - **fio --rw**=<opt1> **--bs**==<opt2> **--size**=<opt3> **--filename**=<opt4> **--direct**=<opt5> **--ioengine**=libaio **--name**=isotdaq

```
[student@storage_lecture]$ fio --rw=write --bs=1M --size=1G --filename=deleteme
--direct=0 --ioengine=libaio --name=isotdaq

fio-3.12
Starting 1 process
isotdaq : Laying out IO file (1 file / 1024MiB)
… … …
Run status group 0 (all jobs):

  WRITE: bw=276MiB/s (282MB/s), 276MiB/s-276MiB/s (282MB/s-282MB/s), io=1024MiB
(1074MB), run=4424-4424msec
```

Redundant Array of Inexpensive Disks (**RAID**)

Redundancy and fault tolerance

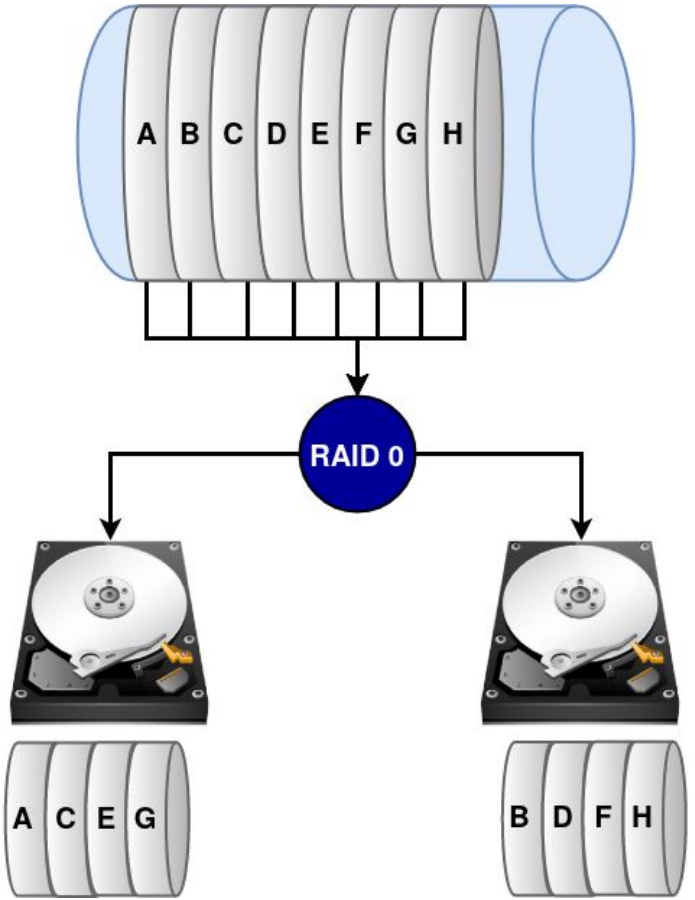
- Multiple physical disk drives are logically grouped into one or more units to increase data performance and/or data redundancy
- Invented in 1987 by researchers from the University of California
- Most common RAID types: RAID 0, RAID 1, RAID 5, RAID 10
- **Fault tolerance** guaranteed by using **parity** as an error protection scheme
  - Based on the XOR logic operation
  - For series of XOR operations, count the number of occurrences of 1:
    - If result is <u>even</u> then bit parity is 0
    - If result is <u>odd</u> then bit parity is 1

A	B	A XOR B
0	0	0
0	1	1
1	0	1
1	1	0

Redundant Array of Inexpensive Disks (**RAID**)

RAID 0 - Striping

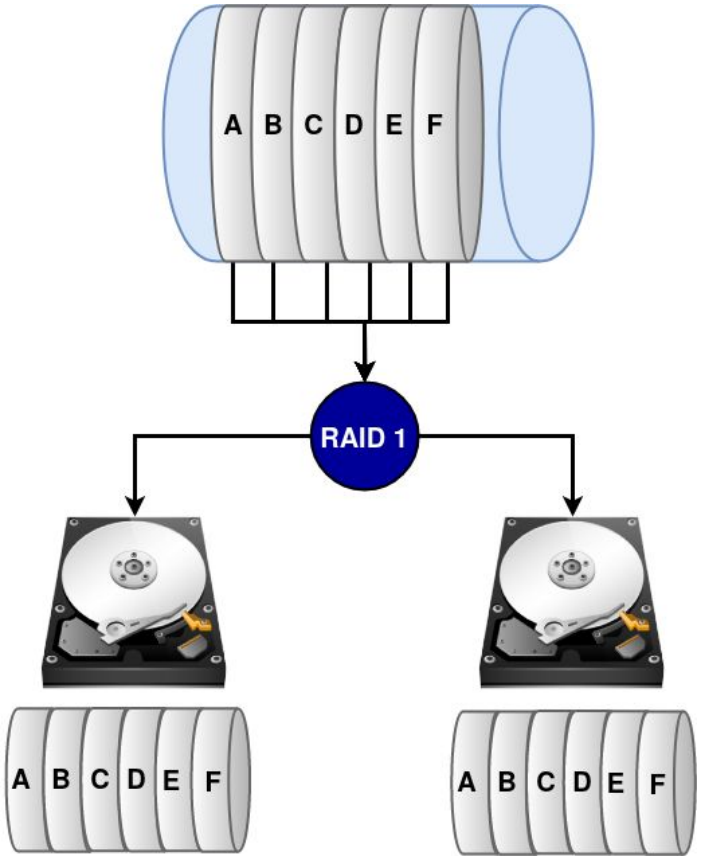
- Data divided in blocks and striped across multiple disks
- **Not fault tolerant** because data is not duplicated
- Speed advantage
  - Two disk controllers allow to access data much faster



Redundant Array of Inexpensive Disks (**RAID**)

RAID 1 - Mirroring and Duplexing

- Data divided in blocks and copied across multiple disks
- **Fault tolerant** because of data mirroring
    - Each disk has the same data
- **Disadvantage**: usable capacity is half of the total



Redundant Array of Inexpensive Disks (**RAID**)

Redundancy and fault tolerance

- Multiple physical disk drives are logically grouped into one or more units to increase data performance and/or data redundancy
- Invented in 1987 by researchers from the University of California
- Most common RAID types: RAID 0, RAID 1, RAID 5, RAID 10
- **Fault tolerance** guaranteed by using **parity** as an error protection scheme
    - Based on the XOR logic operation
    - For series of XOR operations, count the number of occurrences of 1:
        - If result is <u>even</u> then bit parity is 0
        - If result is <u>odd</u> then bit parity is 1

A	B	A XOR B
0	0	0
0	1	1
1	0	1
1	1	0

A crash course on bit parity

Example for a “3-bit” hard drive

Disk 1	Disk 2	Disk 3	Count	Parity
0	1	1		
1	0	0		
1	1	0		

A crash course on bit parity

Example for a “3-bit” hard drive

Disk 1	Disk 2	Disk 3	Count	Parity
0	1	1	2	0
1	0	0	1	1
1	1	0	2	0

A crash course on bit parity

Disk failure

Disk 1	Disk 2	Disk 3	Count	Parity
0	1	1	2	0
1	0	0	1	1
1	1	0	2	0

A crash course on bit parity

Example for a “3-bit” hard drive

Disk 1	Disk 2	Parity	Count	Disk 3
0	1	0		
1	0	1		
1	1	0		

A crash course on bit parity

Example for a “3-bit” hard drive

Disk 1	Disk 2	Parity	Count	Disk 3
0	1	0	1	1
1	0	1	2	0
1	1	0	2	0

A crash course on bit parity

Example for a “3-bit” hard drive

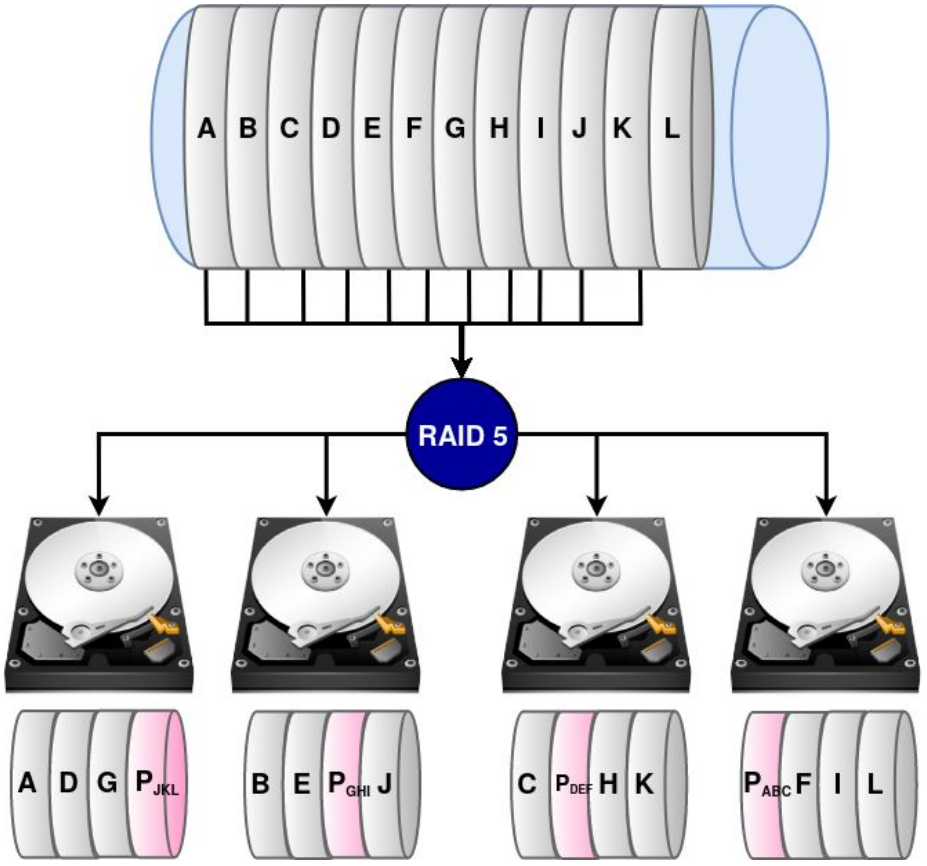
Disk 1	Disk 2	Parity	Count	Disk 3
0	1	0	1	1
1	0	1	2	0
1	1	0	2	0

Disk 3
1
0
0

Redundant Array of Inexpensive Disks (**RAID**)

RAID 5 - Striping with parity

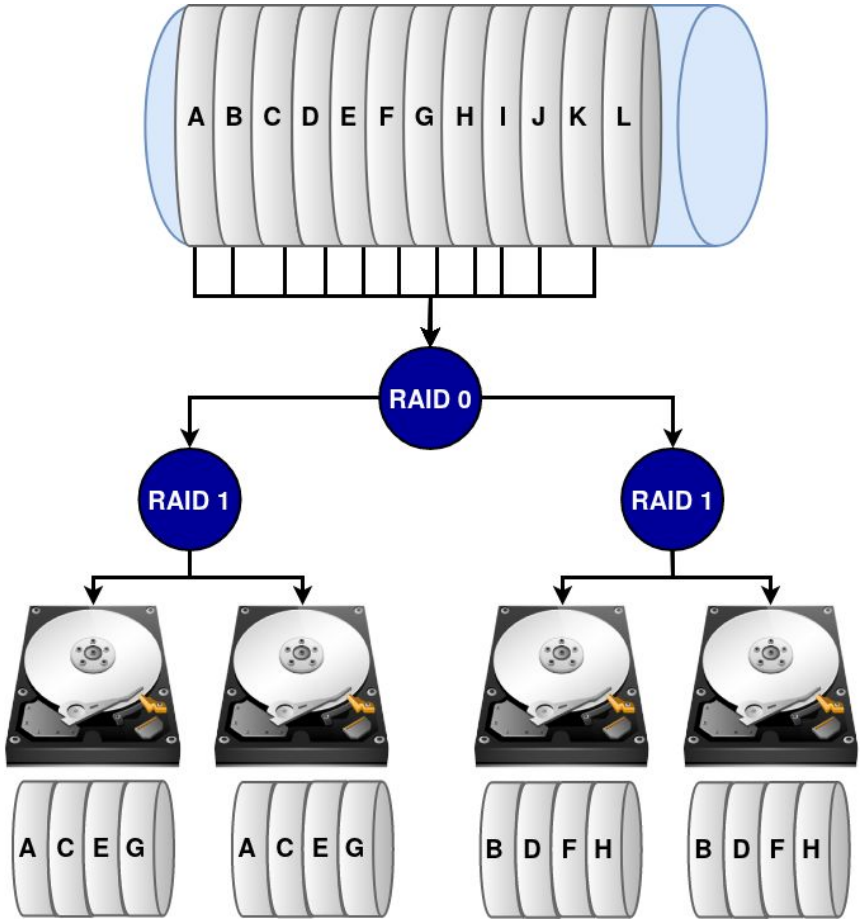
- Requires 3 or more disks
- Data is not duplicated but **striped** across multiple disks
- Fault tolerant because **parity** is also striped with the data blocks
- Larger capacity provided compared to RAID 1
- Disadvantage: an entire disk is used to store parity



Redundant Array of Inexpensive Disks (**RAID**)

RAID 10 = RAID 1 + RAID 0

- Requires a minimum of 4 disks
- Data is **striped** (RAID 0)
- Data is duplicated across multiple disks (RAID 1)
- **Advantage**: fault tolerance and higher speed
- **Disadvantage**: only half of the available capacity is usable



Redundant Array of Inexpensive Disks (**RAID**)

HW, SW

- **Hardware** implementation:
  - Use of RAID controllers
  - Manage system independently of OS
  - Offload I/O operation and parity computation
  - Cost usually high
- **Software** implementation:
  - OS used to manage RAID configuration
  - Impact on CPU usage can be high
- **Disadvantage**: scaling to multiple servers is not possible

Redundant Array of Inexpensive Disks (**RAID**)

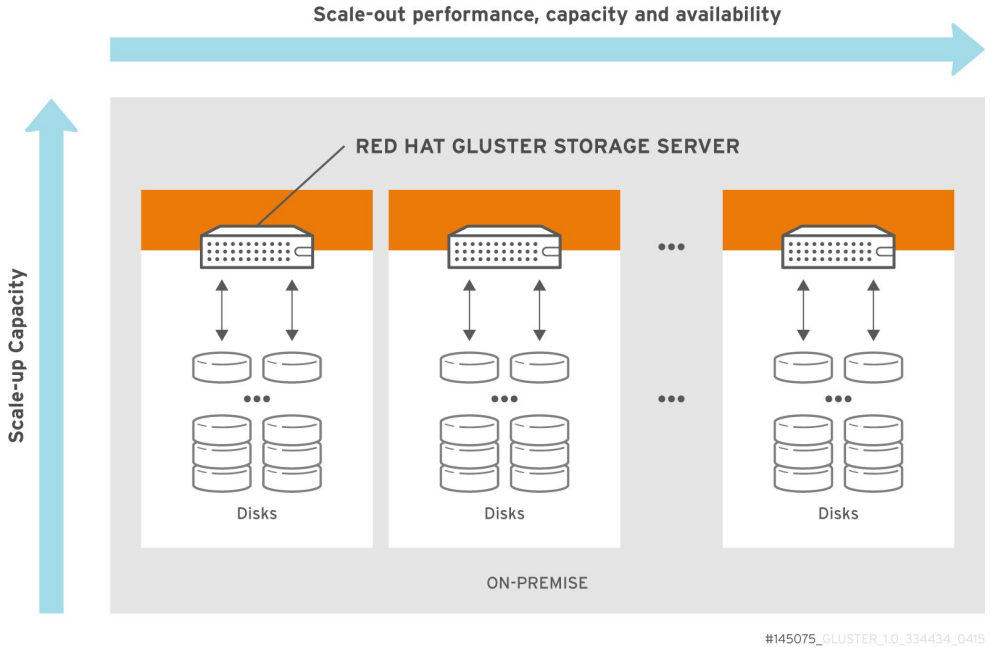
HW, SW

- **Hardware** implementation:
  - Use of RAID controllers
  - Manage system independently of OS
  - Offload I/O operation and parity computation
  - Cost usually high
- **Software** implementation:
  - OS used to manage RAID configuration
  - Impact on CPU usage can be high
- **Disadvantage**: scaling to multiple servers is not possible

Distributed storage systems

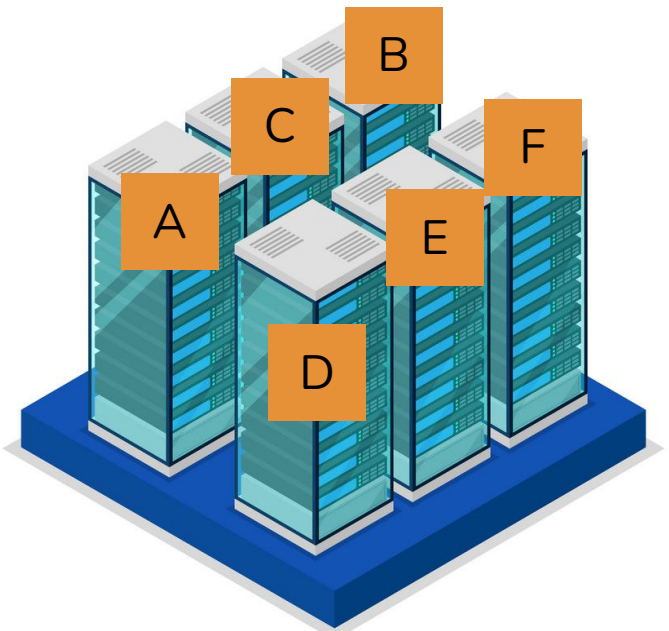
Distributed storage systems

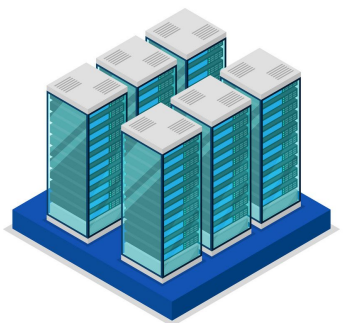
- **Distributed storage system**: files are shared and distributed between multiple nodes
  - Active communities (Red Hat, IBM, Apache)
  - Example: Ceph, Gluster, Hadoop, Lustre
  - Used by some experiments (CMS)
  - Interesting features:
    - load balancing
    - data replication
    - smart placement policies
    - scaling up to O(1000) nodes



Distributed storage systems

- **Application in DAQ**: implementation of the **event builder**:
  - **Physical event building** (**traditional approach**): data fragments are fetched explicitly over a network from temporary buffers at the readout nodes to a single physical location

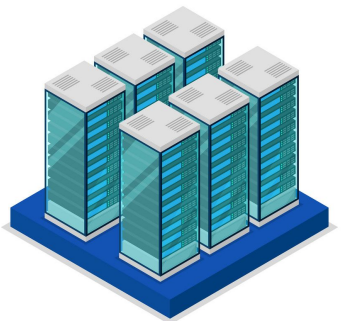




Distributed storage systems

- **Application in DAQ**: implementation of the **event builder**:
  - **Physical event building** (**traditional approach**): data fragments are fetched explicitly over a network from temporary buffers at the readout nodes to a single physical location

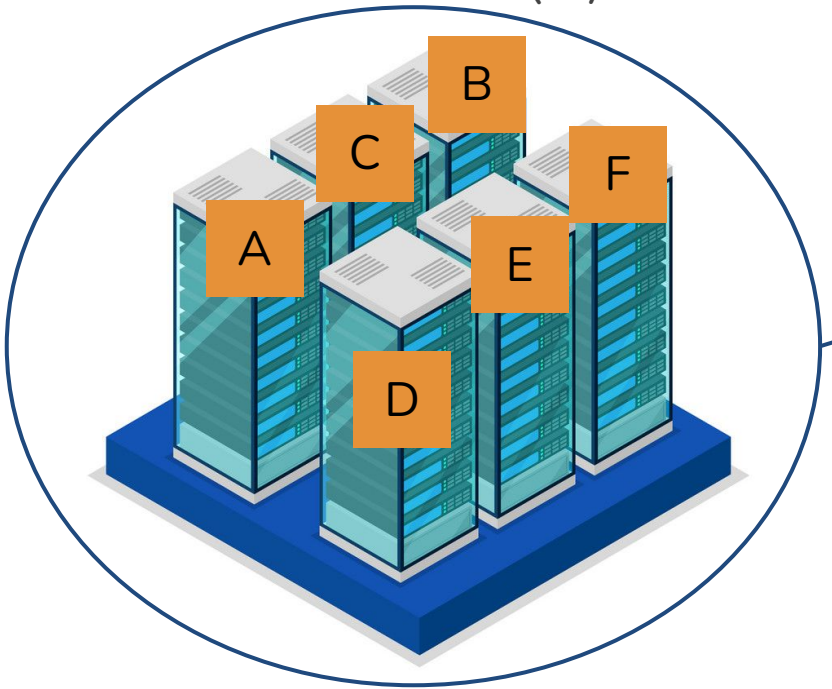
Network & copies



A B C D E F

Distributed storage systems

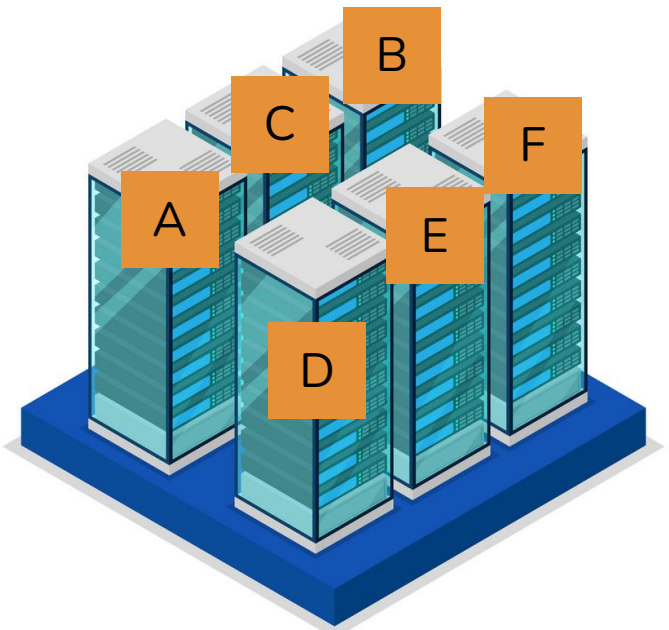
- **Application in DAQ**: implementation of the **event builder**:
  - **Logical event building**: fragments are stored in a large distributed system and events are built by computing the location of the fragments (metadata operation)
- **R&D** for future DAQ systems: ATLAS (Phase-II), DUNE, etc.



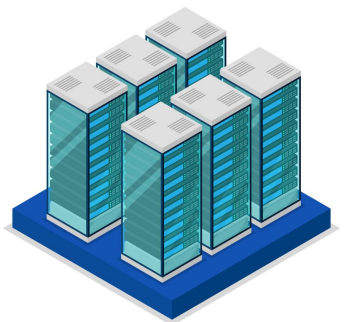
Intel DAOS
(Distributed Asynchronous Object Store)

Distributed storage systems

- **Application in DAQ**: implementation of the **event builder**:
  - **Logical event building**: fragments are stored in a large distributed system and events are built by computing the location of the fragments (metadata operation)
- **R&D** for future DAQ systems: ATLAS (Phase-II), DUNE, etc.



Fragment addresses



&A &B &B &C &D &E &F

DAQ takeaway

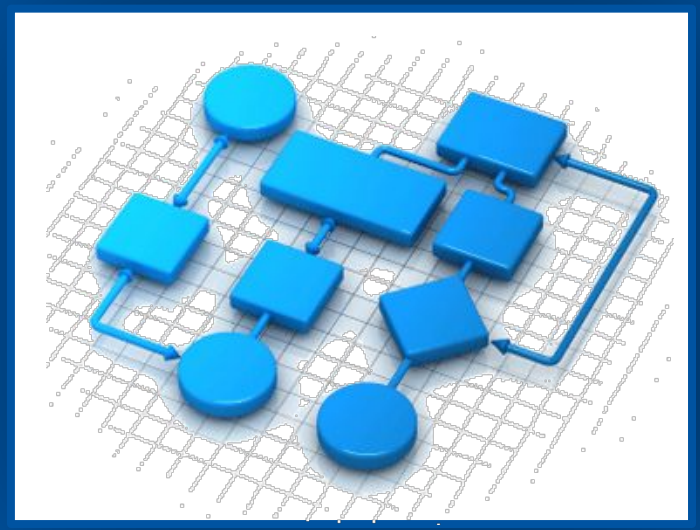
Storage technologies

- Different storage media available on the market for different use cases
  - Long term storage, mostly sequential access → HDD
  - Low latency and large capacity → SSD
  - High rate and persistent → Non-Volatile memory
  - Fast and temporary → DRAM
- Keep in mind that **price/GB** changes a lot for different storage media
- When designing a DAQ system always keep an eye on the target throughput and required rate for your application
- **Data safety** and **reliability** is an important factor!
  - RAID

DAQ takeaway

Storage challenges for the next generation DAQ systems

- Physics signals are rare!
  - Higher intensity beams are needed
  - More granular detectors
  - Consequence: store more data
- HL-LHC: Data rates and data bandwidths will increase by ~ 1 order of magnitude
  - Consequence: scale DAQ system
  - Use commercial off-the-shelf technology as much as possible
- Current storage landscape
  - HDD: large and cheap streaming storage
  - SSD: low latency and high throughput



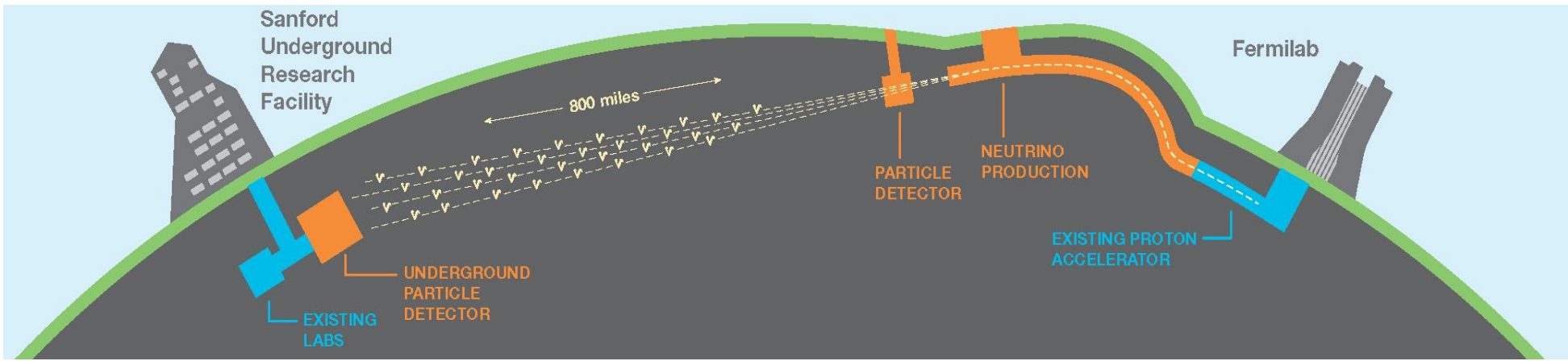
Storage systems in HEP

DEEP UNDERGROUND NEUTRINO EXPERIMENT

DUNE experiment

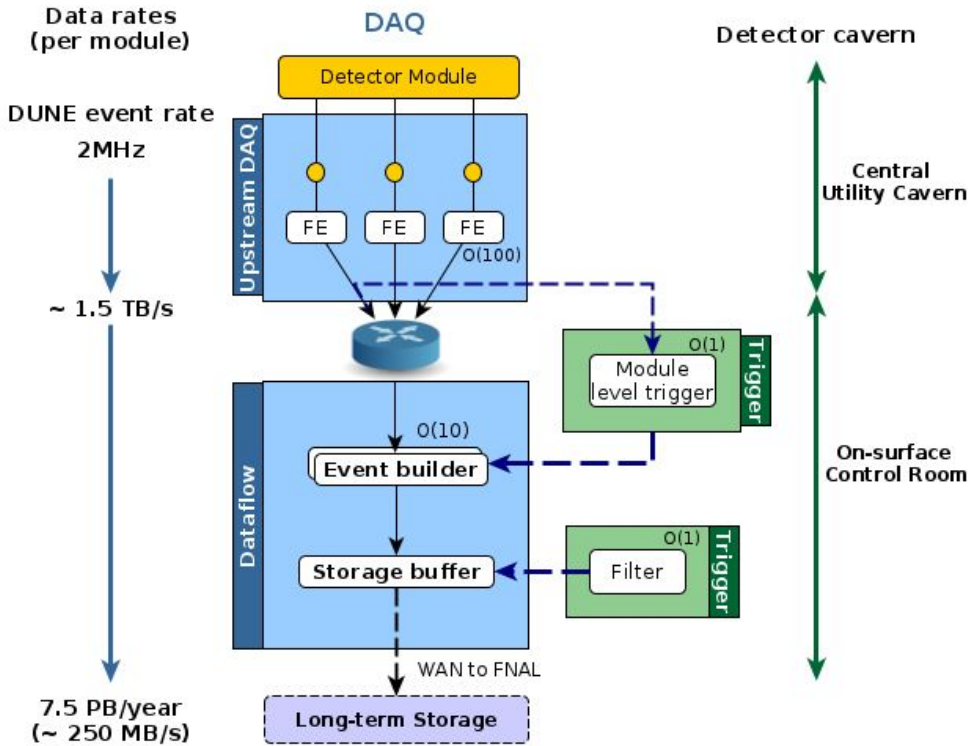
Quick overview

- Neutrino experiment located at Sanford Underground Research Facility in South Dakota
- Far detector located 1300 km away from source and approximately 1.48 km underground
- 4 modules of 17 kton LAr time projection chamber



DUNE experiment

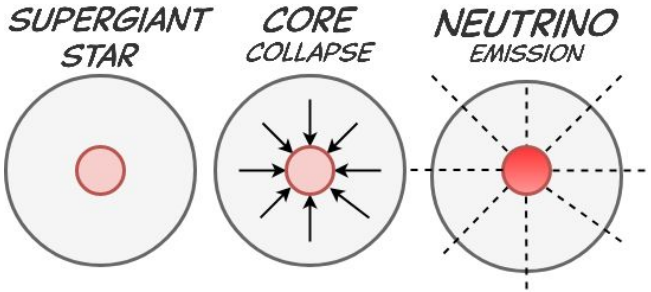
DAQ system



- TPC sampling rate: 2 MHz
- Each readout board :
  - 10 links
  - O(1) GB/s per link

  } 10 GB/s
- 150 detector units
  - Total readout rate O(1.5) TB/s

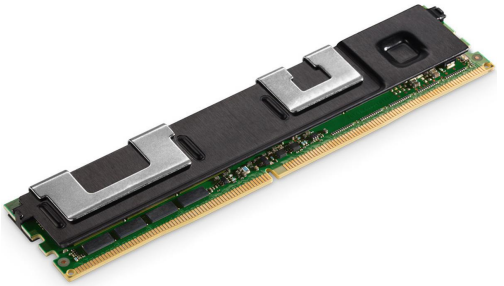
Supernova Neutrino Burst



- **Supernova Neutrino Burst** (SNB) detection
  - One of the physics goals of DUNE
  - Detection of **rare**, **low energy** and **distributed** signatures
- Data taking of SNB events is **complex**:
  - Long trigger latency
  - Physics event distributed over time
  - Critical data: avoid any potential loss
- **Requirements**:
  - Transient buffer O(10) seconds (i.e. 15 TB per detector module)
  - On trigger: persist O(100) seconds (i.e. 150 TB per detector module)

Supernova Neutrino buffer

Persistent memory



- Critical data and high bandwidth:
  - Use of Non-Volatile Memory technology (3D XPoint)
- **Successful** prototype capable of buffering data from the readout system
  - Transient buffer of 10 seconds
  - Store for over 100 seconds
  - Sustained a maximum throughput of 10 GB/s
- From benchmark results: the bandwidth of NVM is approximately 10 GB/s
- Successfully integrated in DUNE DAQ software

```
00003040  e3 a9 35 8e 66 92 63 e8  39 8d 70 4d e8 69 8a 93  |..5.f.c.9.pM.i..|
00003050  6f dc 29 d8 94 8f f7 b1  98 b3 92 3a 24 a3 d9 b3  |o.).........:$...|
00003060  91 3c 27 c0 99 53 97 3f  c3 69 c3 39 3a 99 9c d4  |.<'..S.?.i.9:...|
00003070  23 f8 3b 89 c0 bc c3 b8  3a 8d 33 ab 29 18 90 92  |#.;.....:.3.)...|
00003080  00 00 5f 30 0f ee 20 46  00 00 00 00 aa aa aa aa  |.._0.. F........|
00003090  42 0d c9 39 8d 93 36 bf  59 53 91 3d 27 ec 49 f3  |B..9..6.YS.='.I.|
000030a0  97 3c 71 e3 79 63 a5 40  9f a0 f3 09 37 97 94 e6  |.<q.yc.@....7...|
000030b0  e3 78 37 8e 92 fe 73 f8  39 8d 5b 2a 99 69 8d 96  |.x7...s.9.[*.i..|
000030c0  e7 bf 88 08 8d 9c c2 b3  88 03 8f 39 18 a1 69 e3  |...........9..i.|
000030d0  92 38 ff af 18 b3 98 3e  89 74 f3 a9 37 8e 81 91  |.8.....>.t..7...|
000030e0  d3 f9 39 8d ab 5c 73 89  37 8e 84 cc d8 69 8e 91  |..9..\s.7....i..|
000030f0  00 00 38 50 9d 75 20 46  00 00 00 00 aa aa aa aa  |..8P.u F........|
00003100  26 f0 49 48 8d 87 45 4e  e9 c3 8b 38 a4 84 88 93  |&.IH..EN...8....|
00003110  91 39 76 98 99 a3 95 41  c0 9a 33 18 3c 8f db 58  |.9v....A..3.<..X|
00003120  83 08 3b 85 82 5e 43 78  3a 87 da 31 98 e8 8b 85  |..;..^Cx:..1....|
00003130  e4 f7 d8 38 90 8f b8 9e  f8 23 8d 3d 54 a1 19 63  |...8.....#.=T..c|
00003140  93 39 40 be 29 d3 99 3f  ca 3c 23 e9 37 8f 78 dc  |.9@.)..?.<#.7.x.|
00003150  c3 48 38 8d 76 a4 f3 48  37 8e a1 7f 38 48 8d 8f  |.H8.v..H7...8H..|
00003160  00 00 11 8e 88 26 20 46  00 00 00 00 aa aa aa aa  |.....& F........|
00003170  98 6b f8 29 89 95 1b fd  79 13 8d 3e c1 ba 18 b3  |.k.)....y..>....|
00003180  8f 3f 41 07 39 04 94 42  b3 85 33 89 38 97 8d 9d  |.?A.9..B..3.8...|
00003190  53 49 38 9e 95 b7 23 39  38 9a f6 4b 88 d9 89 93  |SI8...#98..K....|
000031a0  1f d8 c9 38 94 92 33 b8  c9 33 93 3a 8d ad c9 d3  |...8..3..3.:....|
000031b0  09 3c c5 df 79 a3 a2 41  d1 d3 33 58 3a 8f a6 5b  |.<..y..A..3X:..[|
```

Conclusions

- DAQ mentoring:
  - Storage system is crucial for physics results
  - Online data taking has different requirements from offline analysis
- Design of a storage system:
  - Focus on both **bandwidth** and **rate**
  - **Latency / access pattern**
  - Several storage media for different use-cases (HDD, SSD, NVM, DRAM)
- Very important to benchmark performance of devices. Tools: DD and FIO
- Use redundancy where necessary based on system availability requirements

ISOTDAQ

International School of Trigger and Data Acquisition

Thank you ! Questions ?

adam.abed.abud@cern.ch
enrico.gamberini@cern.ch