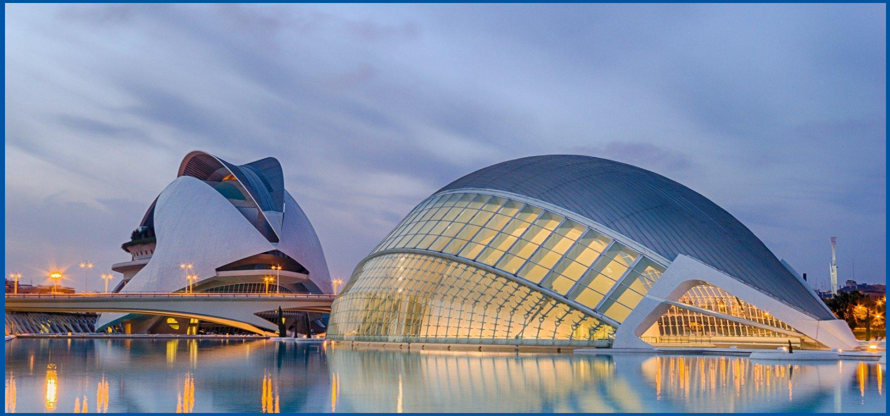
ISOTDAQ

International School of Trigger and Data Acquisition



Storage systems for DAQ

Adam Abed Abud (Univ. of Liverpool / CERN)

ISOTDAQ 2020

January 18, 2020 (Valencia, Spain)

Who stores more data? Instagram or ATLAS experiment?

A: Instagram

B: ATLAS

Who stores more data? Instagram or ATLAS experiment?

A: Instagram

B: ATLAS

Who stores more data? Instagram or ATLAS experiment?

A: Instagram

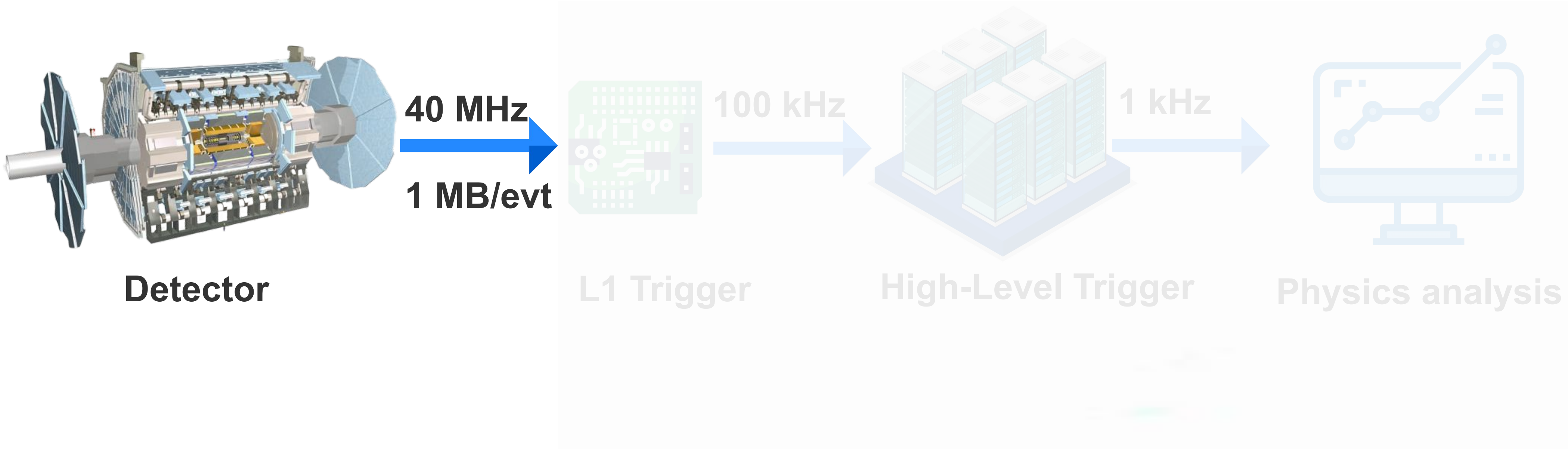
B: ATLAS

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

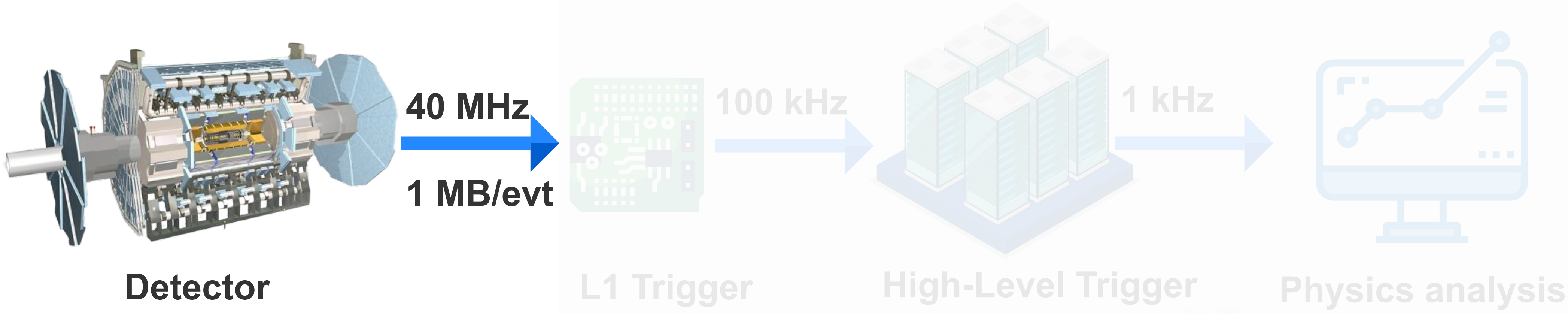
Why are storage systems relevant for DAQ ?

TDAQ pipeline



Why are storage systems relevant for DAQ ?

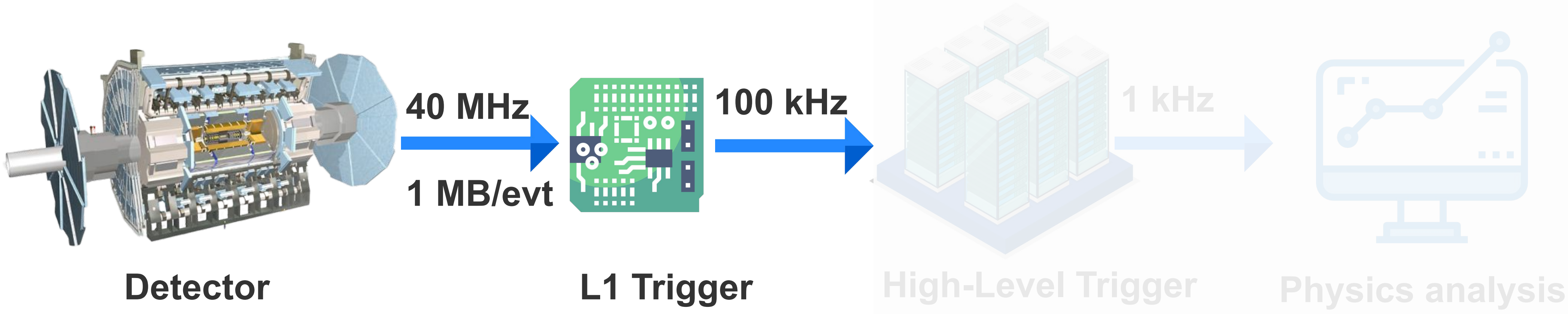
TDAQ pipeline



- All the data produced cannot be stored:
  - Lack of storage resources
  - Not enough (offline) processing power

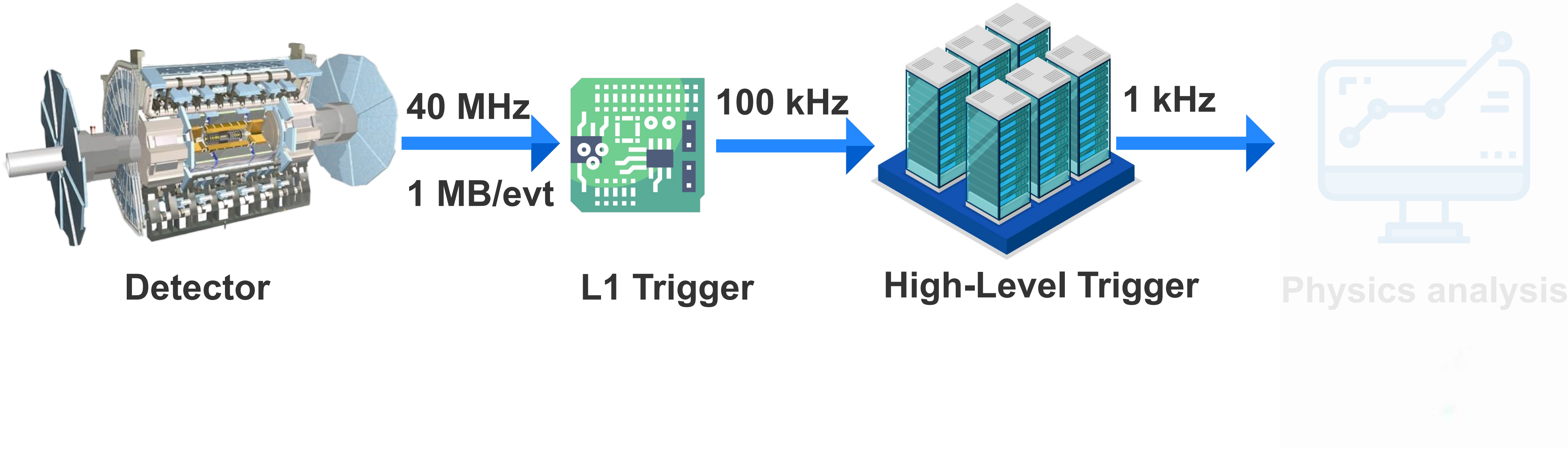
Why are storage systems relevant for DAQ ?

TDAQ pipeline



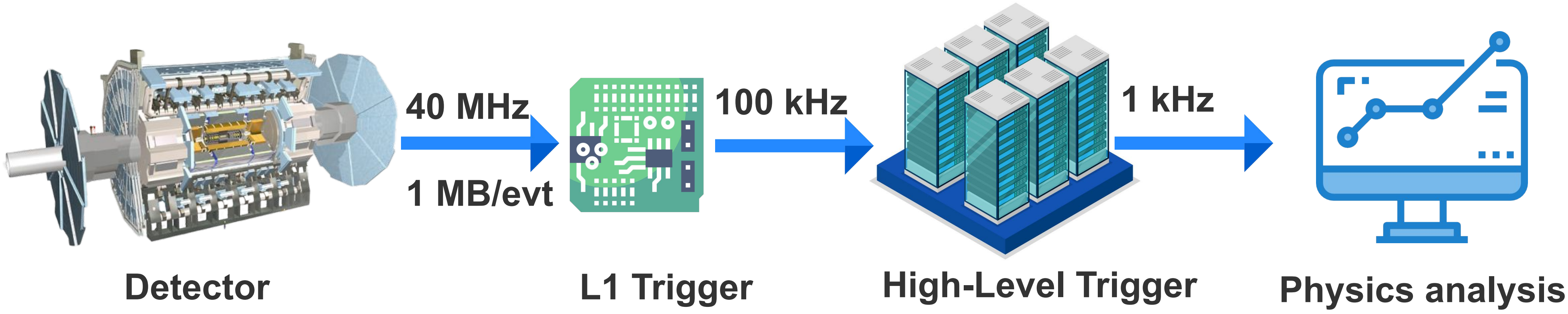
Why are storage systems relevant for DAQ ?

TDAQ pipeline



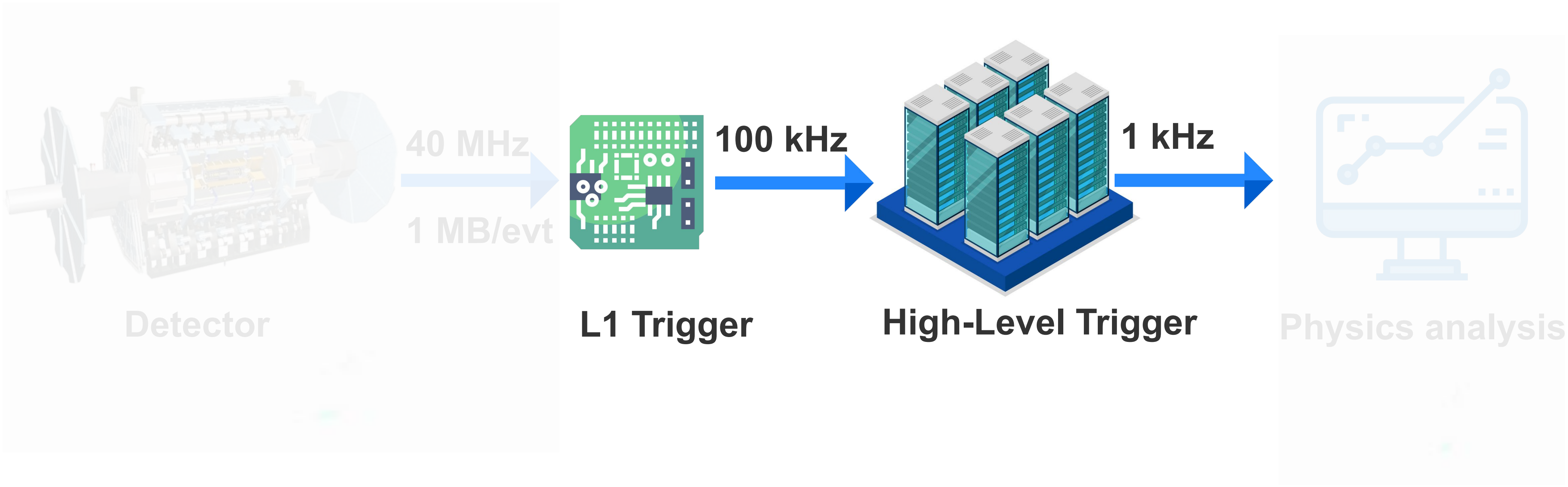
Why are storage systems relevant for DAQ ?

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"

Online vs Offline

- Storage system ensures that data is stored and physics results can be produced!
  - Data stored → physics results
- DAQ requirements different from offline analysis:
  - Storage used to buffer data:
    Absorbs spikes 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**

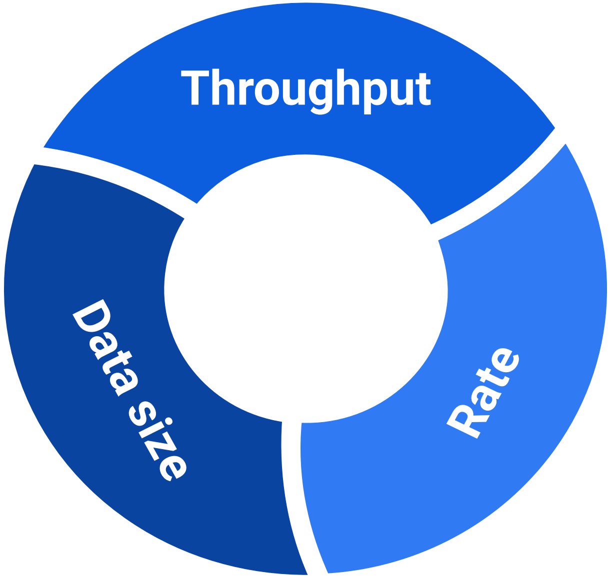
DAQ MENTORING

Online vs Offline

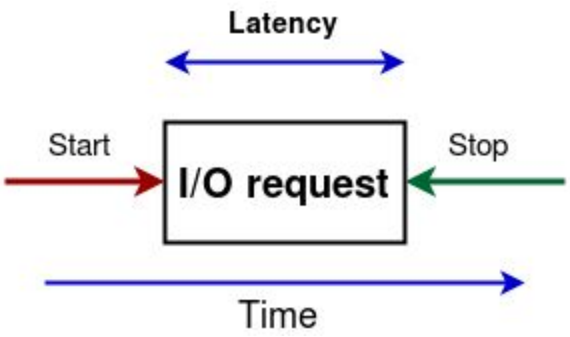
- Storage system ensures that data is stored and physics results can be produced!
  - Data stored → physics results
- DAQ requirements different from offline analysis:
  - Storage used to buffer data:
    Absorbs spikes from the rest of the system
  - Continuous stream of data flow **in and out** the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

DAQ MENTORING
Online vs Offline

- Storage system ensures that data is stored and physics results can be produced!
  - Data stored → physics results
- DAQ requirements different from offline analysis:
  - Storage used to buffer data:
    Absorbs spikes 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**



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

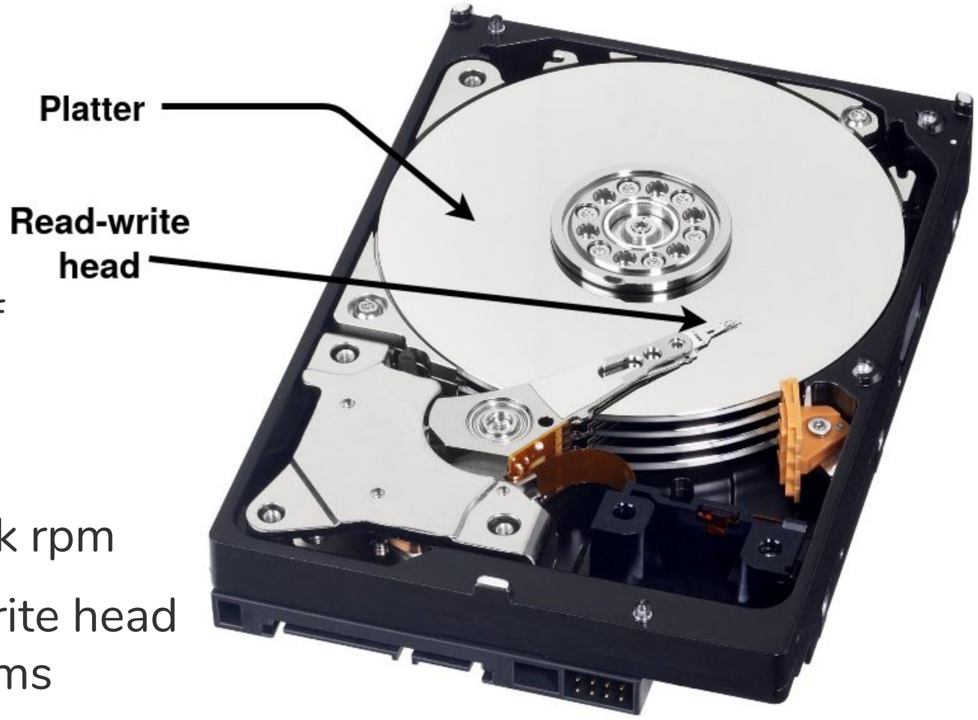
$$\text{Bandwidth} = [\text{I/O block size}] \times [\text{IOPS}]$$

Hard drives (HDD)

Quick introduction

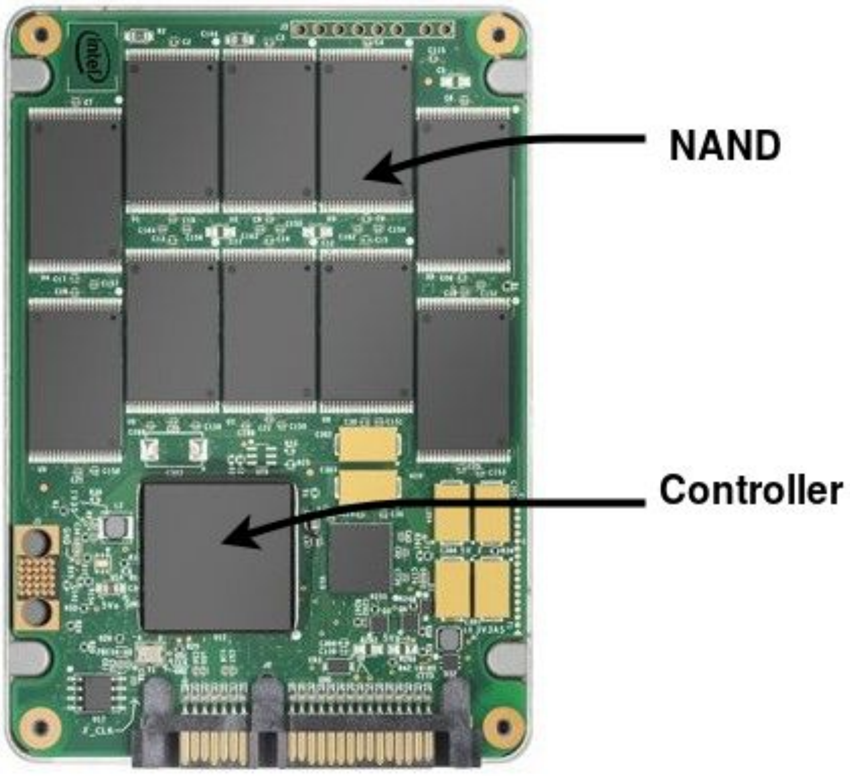
- Electromechanical device
- Circular rotating platter divided into millions of magnetic components where data is stored
- Typical rotational speed of HDDs:
  - 5400 rpm, **7200 rpm**, 10k rpm and 15k rpm
- **Seek time:** time required to adjust the read-write head on the platter. Typical values: from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

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



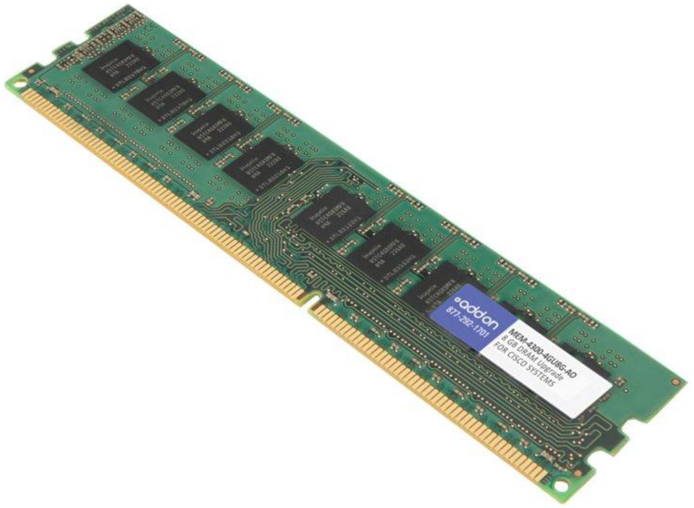
Solid state drives (SSD)
Quick introduction

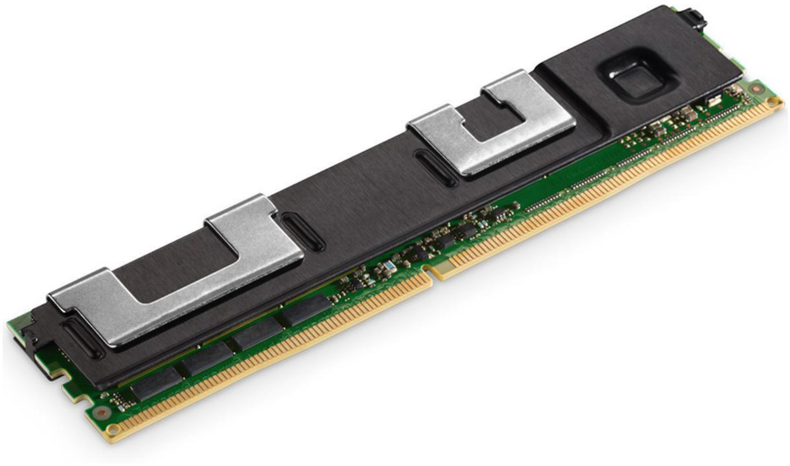
- **Architecture**:
  - NAND flash chipset: store data
  - Controller: caching, load balancing and error handling
- Capacity limited to number of NAND chipsets a manufacturer is able to insert into a device
- (Typically) better performance compared to HDDs
  - There is no mechanical component
  - Reduced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



DRAM and Non-Volatile Memory

Quick introduction

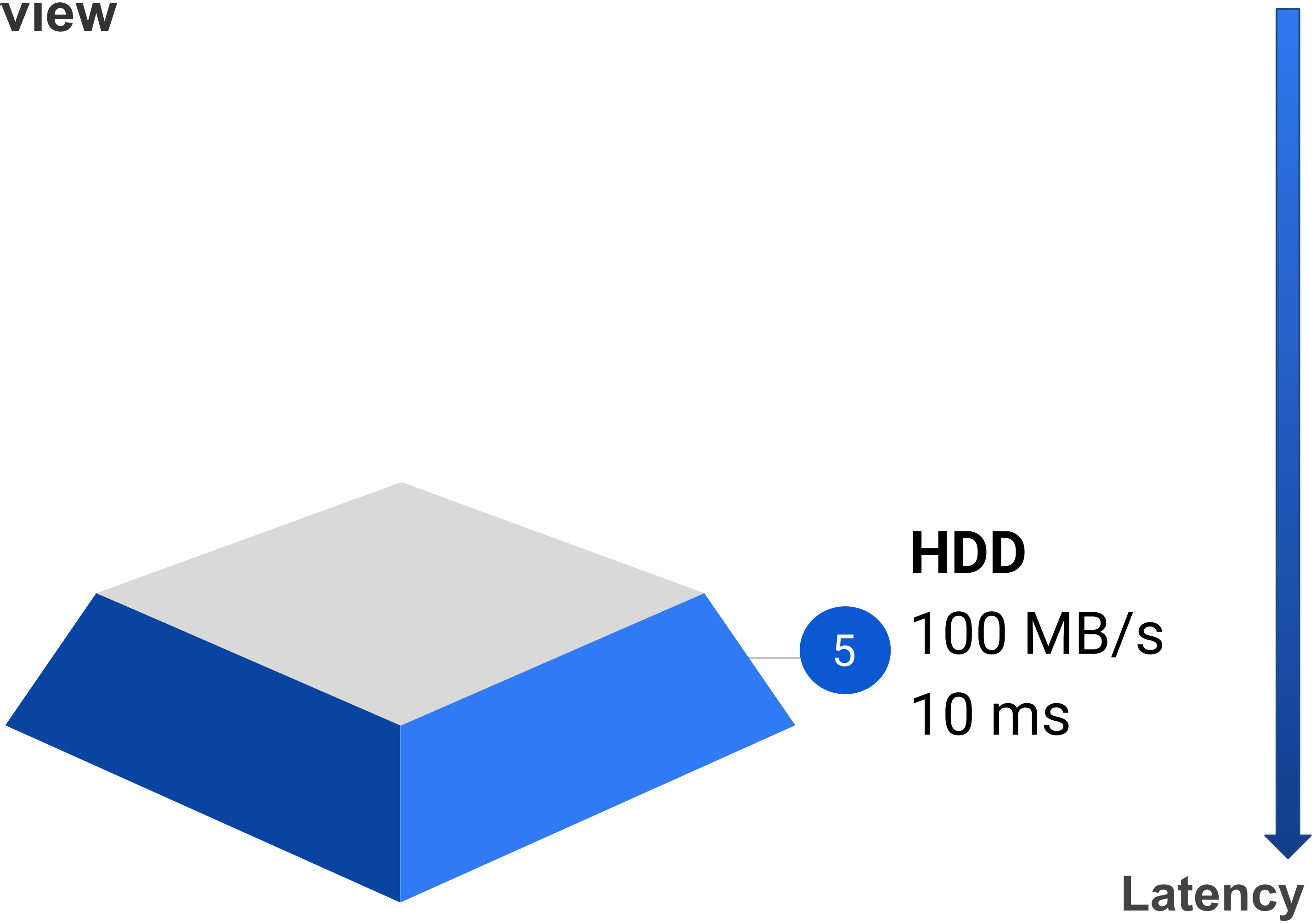




- **DRAM**
  - Semiconductor memory technology
  - Data is not persisted, only temporary storage cells (capacitors and transistors)
  - Low latency (0.1 µs)
- **Non-volatile memory (NVM)**
  - Hold data even if device is turned off
  - Higher storage capacity than DRAM
  - Latency (1 µs)

Latency and Bandwidth

Technology overview



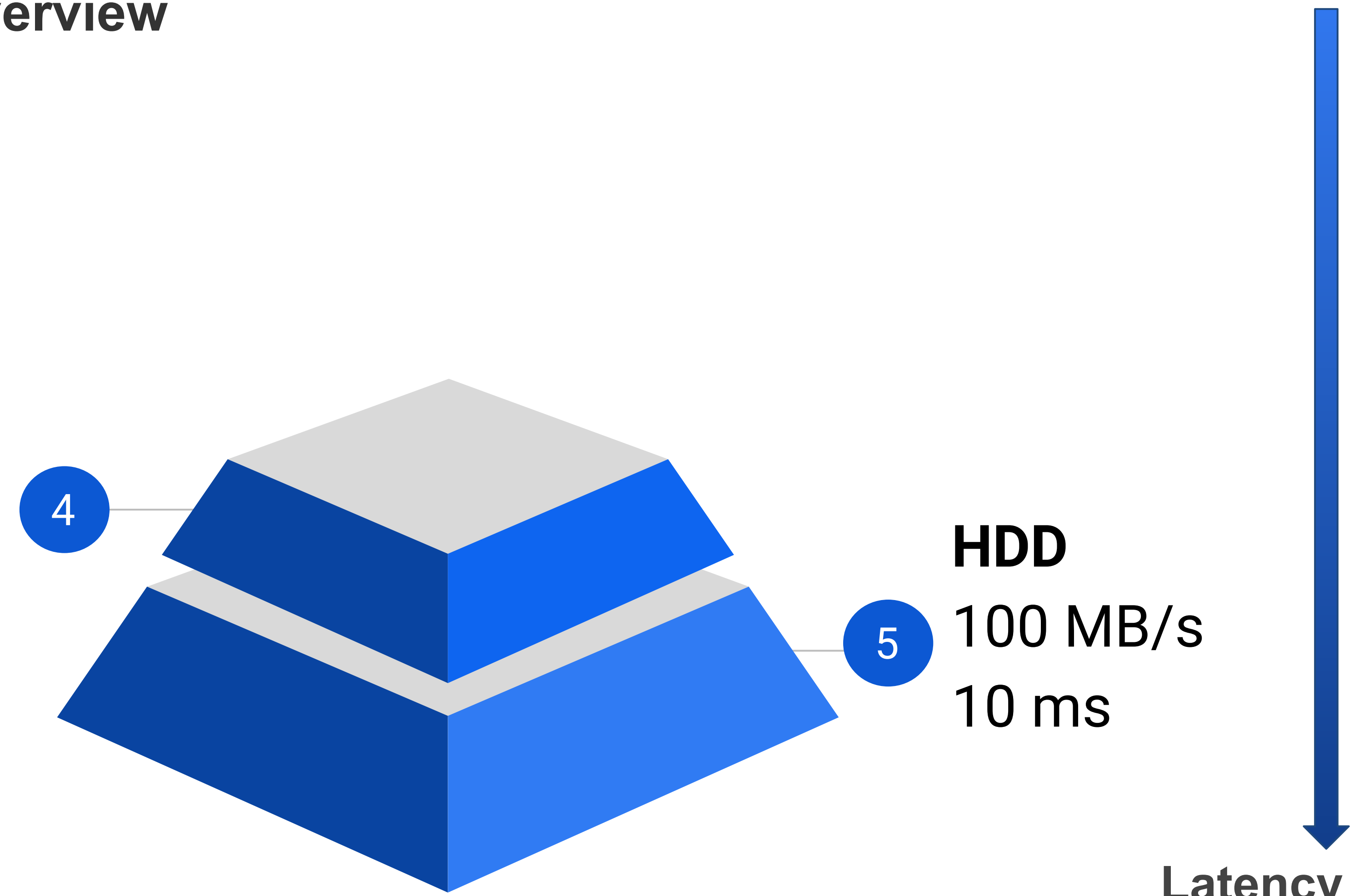
Latency and Bandwidth

Technology overview

SSD

500 MB/s

100 µs



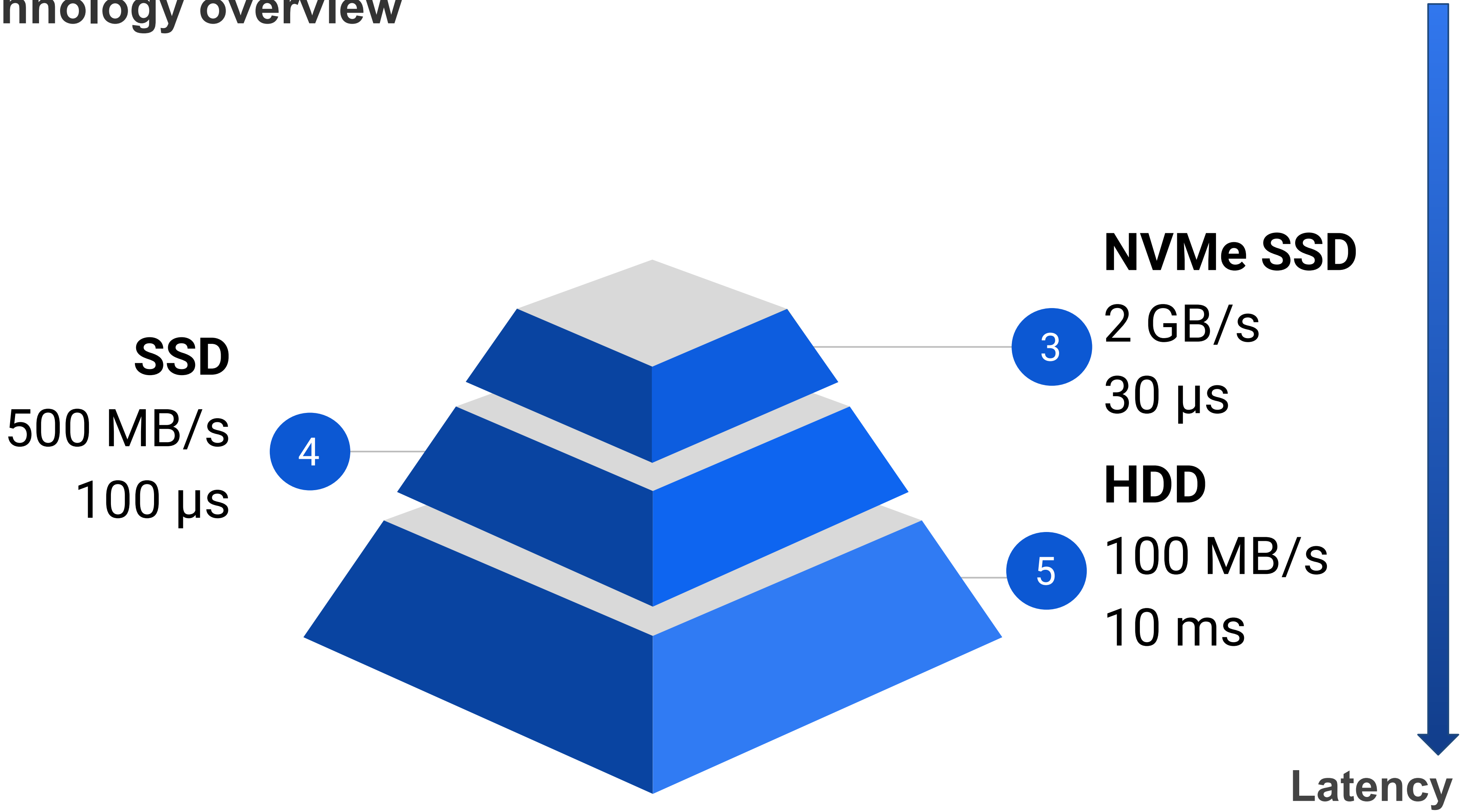
HDD

100 MB/s

10 ms

Latency and Bandwidth

Technology overview



Latency and Bandwidth

Technology overview

NVM

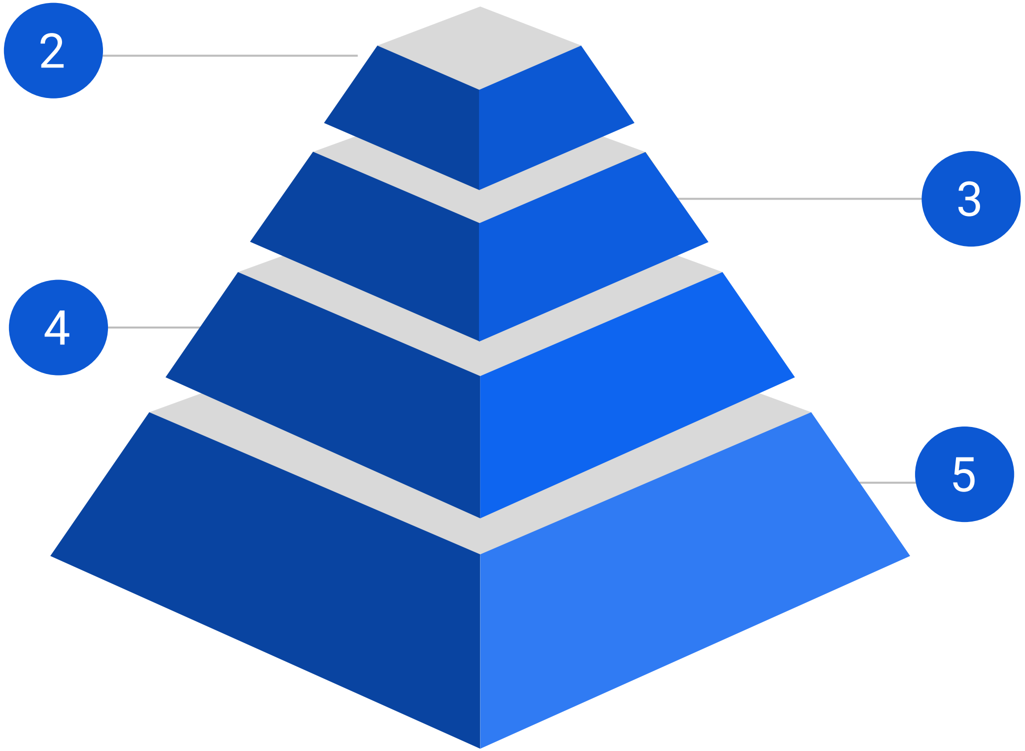
10 GB/s

1 µs

SSD

500 MB/s

100 µs



NVMe SSD

2 GB/s

30 µs

HDD

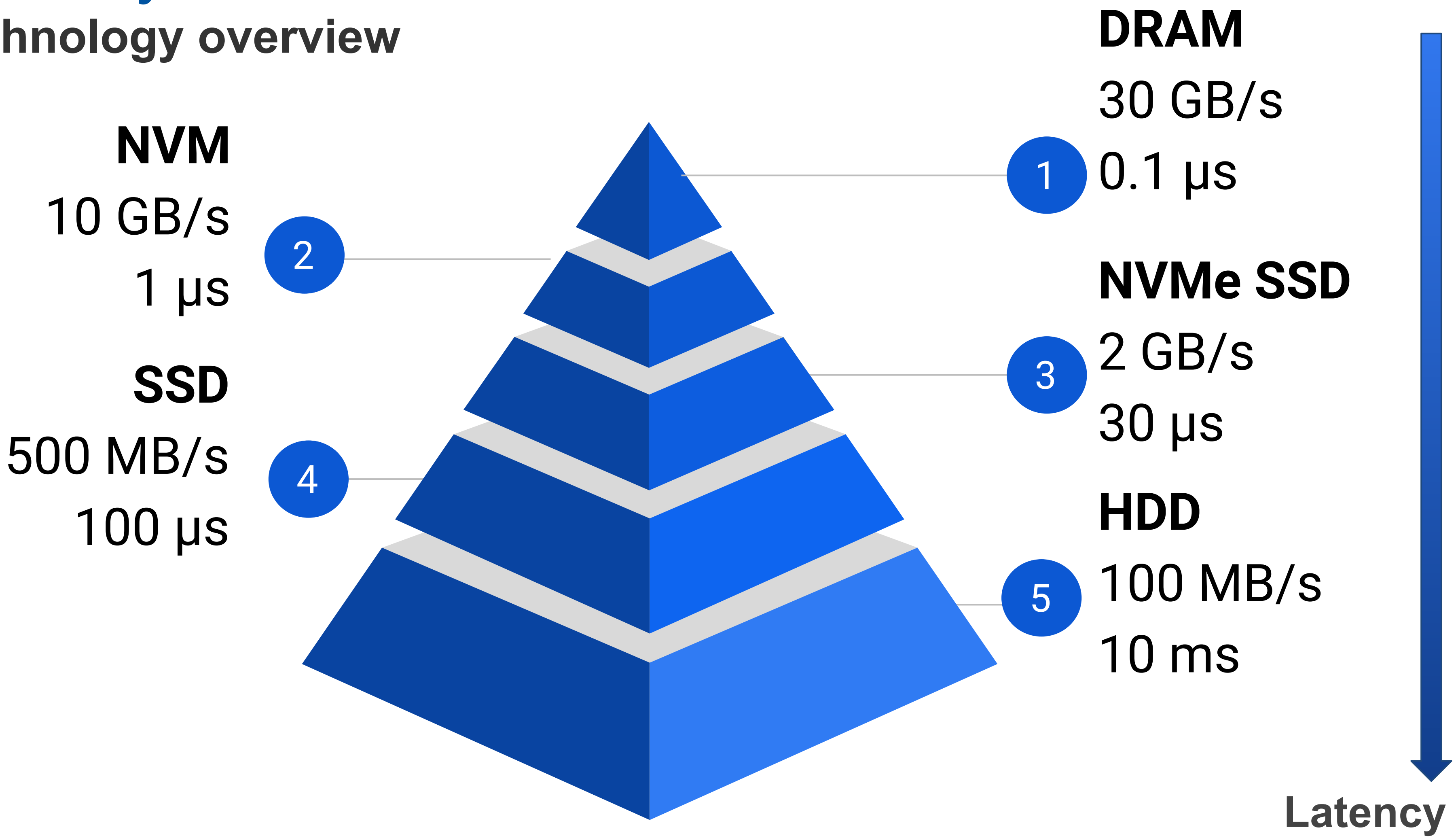
100 MB/s

10 ms

Latency

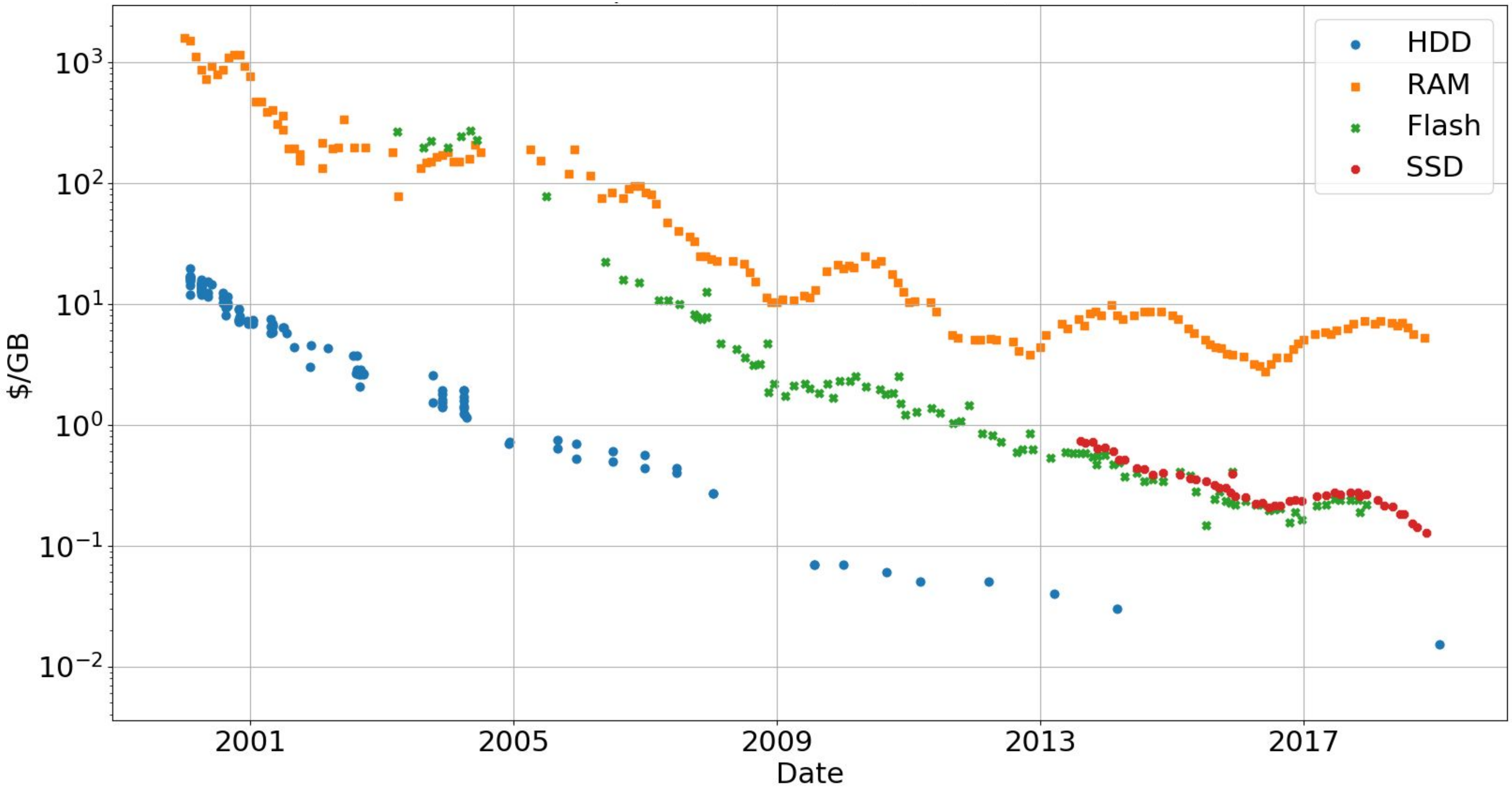
Latency and Bandwidth

Technology overview



Market trend for storage technologies

Price per GB for HDD, SSD, Flash and RAM



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

Redundancy and fault tolerance

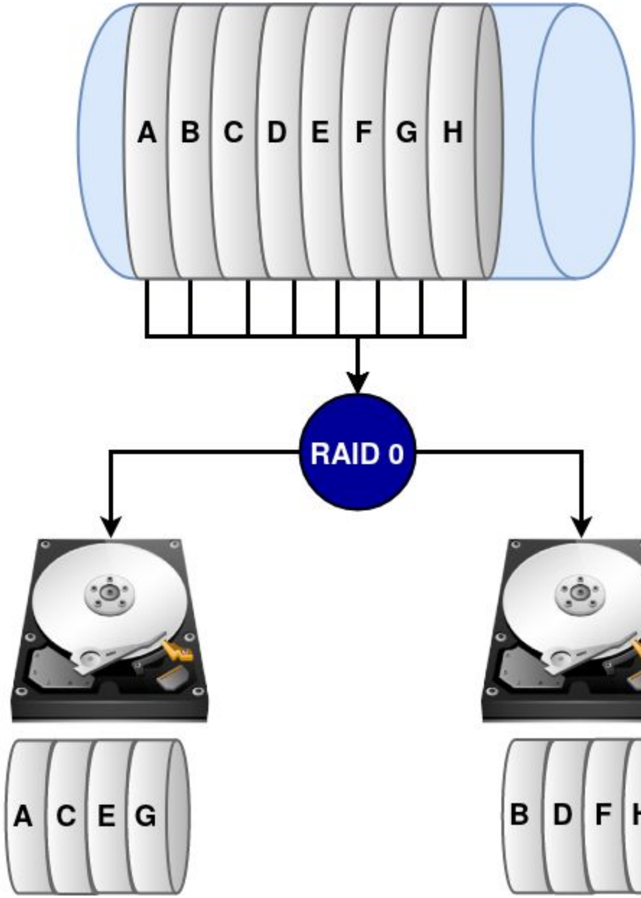
- Multiple physical disk drives are logically grouped into one or more units to increase data performance and/or data redundancy
- Invented in 1987 by researchers from the University of California
- Most common RAID types: RAID 0, RAID 1, RAID 5, RAID 10
- **Fault tolerance** guaranteed by using **parity** 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 even then bit parity is 0
    - If result is odd 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

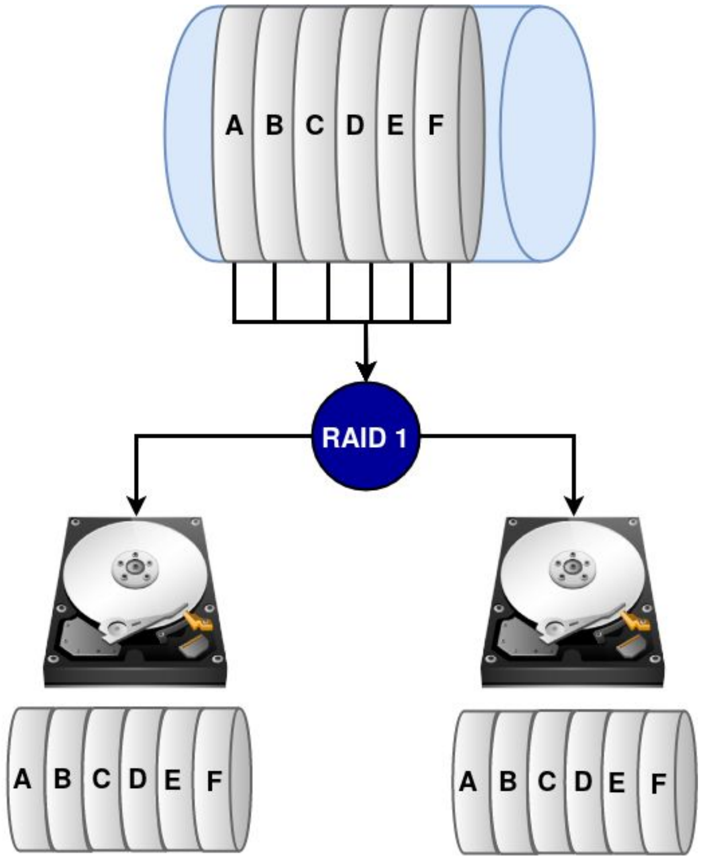
- Data divided in blocks and <u>striped</u> across multiple disks
- **Not fault tolerant** because data is not duplicated
- Speed advantage
    - Two disk controllers allow to access data much faster



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

RAID 1 - Mirroring and Duplexing

- Data divided in blocks and copied across multiple disks
- **Fault tolerant** because 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

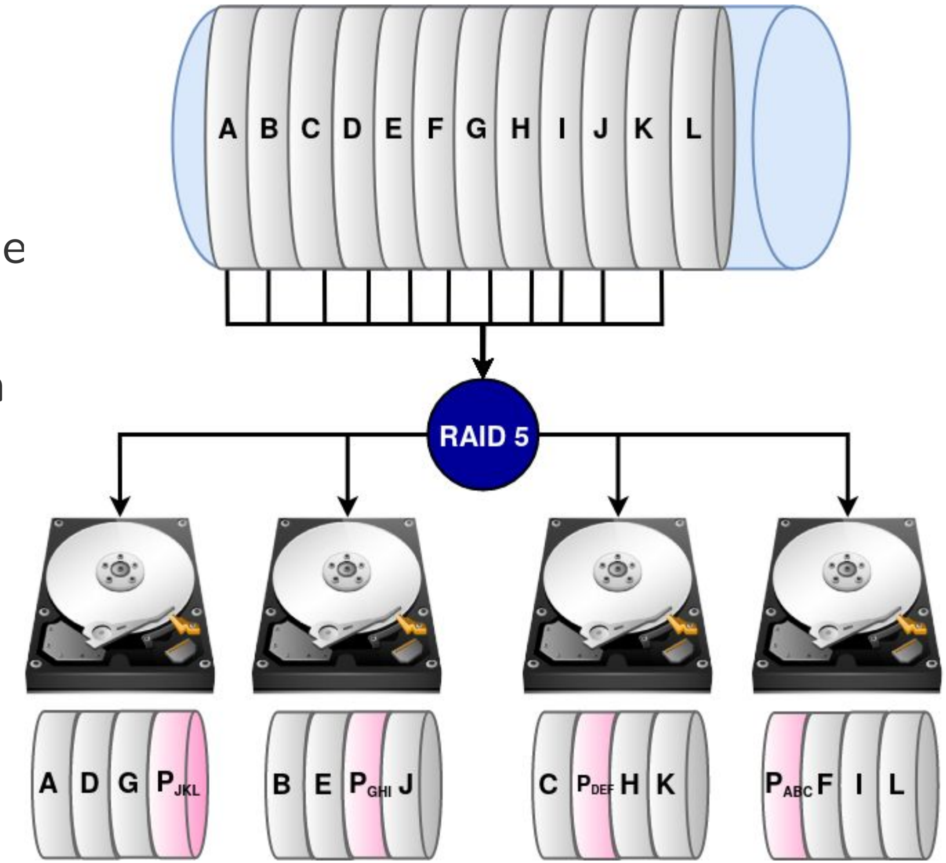
Disk 3
1
0
0

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

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

RAID 5 - Striping with parity

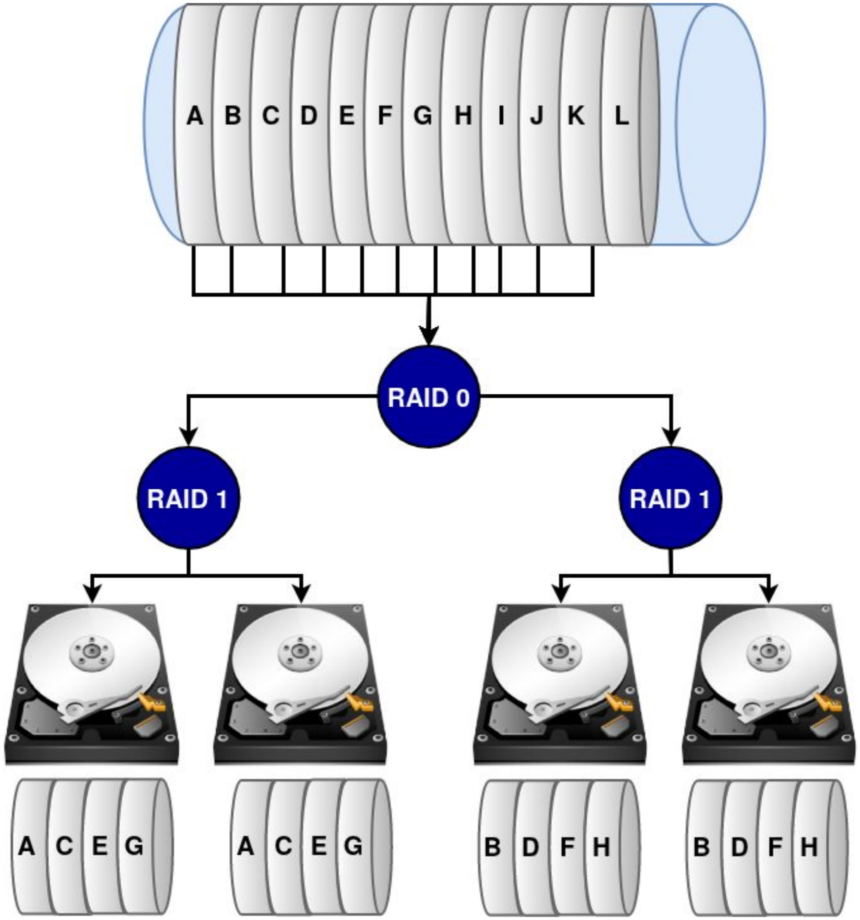
- Requires 3 or more disks
- Data is not duplicated but **striped** across multiple disks
- Fault tolerant because **parity** is also striped with the data blocks
- Larger capacity provided compared to RAID 1
- Disadvantage: an entire disk is used to store parity



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 community (Red Hat, IBM, Apache)
  - Example: Ceph, Gluster, HadoopFS, Lustre
  - Used by some experiments (CMS)
  - Interesting features: load balancing, data replication, smart placement policies, scaling up to O(1000) nodes
- **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
  - **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.

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

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

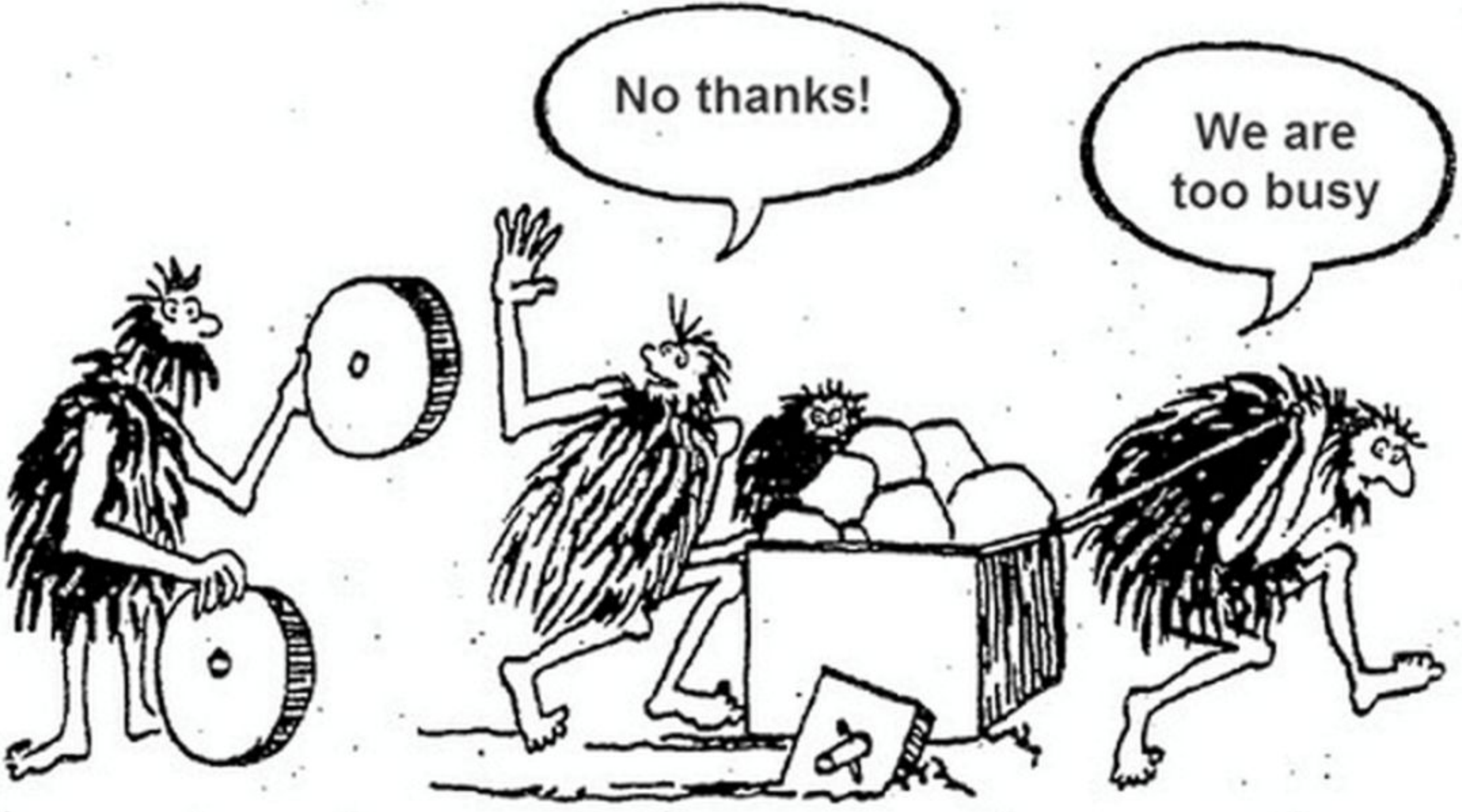
```
[student@storage_lecture]$ fio --rw=write --bs=1M --size=1G --filename=deleteme --direct=0
--ioengine=libaio --name=isotdaq

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
```

Storage challenges for the next generation DAQ systems

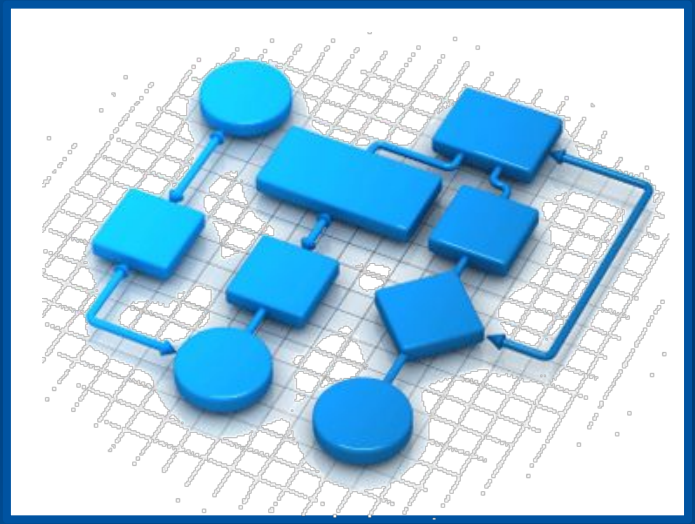
- Physics signals are rare!
    - Higher intensity beams are needed
    - More granular detectors
    - Consequence: store more data
- HL-LHC: Data rates and data bandwidths will increase by ~ 1 order of magnitude
    - Consequence: scale DAQ system
    - Use commercial off-the-shelf technology as much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


No thanks!
We are too busy

DAQ MENTORING

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 storage
  - SSD: low latency and high throughput



Storage systems in HEP

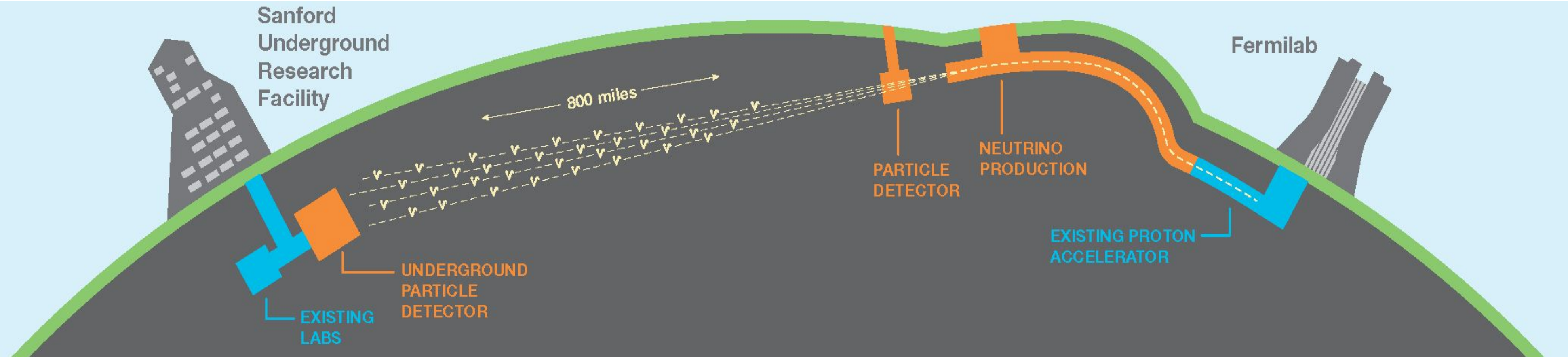
DEEP UNDERGROUND
NEUTRINO EXPERIMENT

DUNE experiment

Quick recap

Talk on DUNE from Alessandro Thea

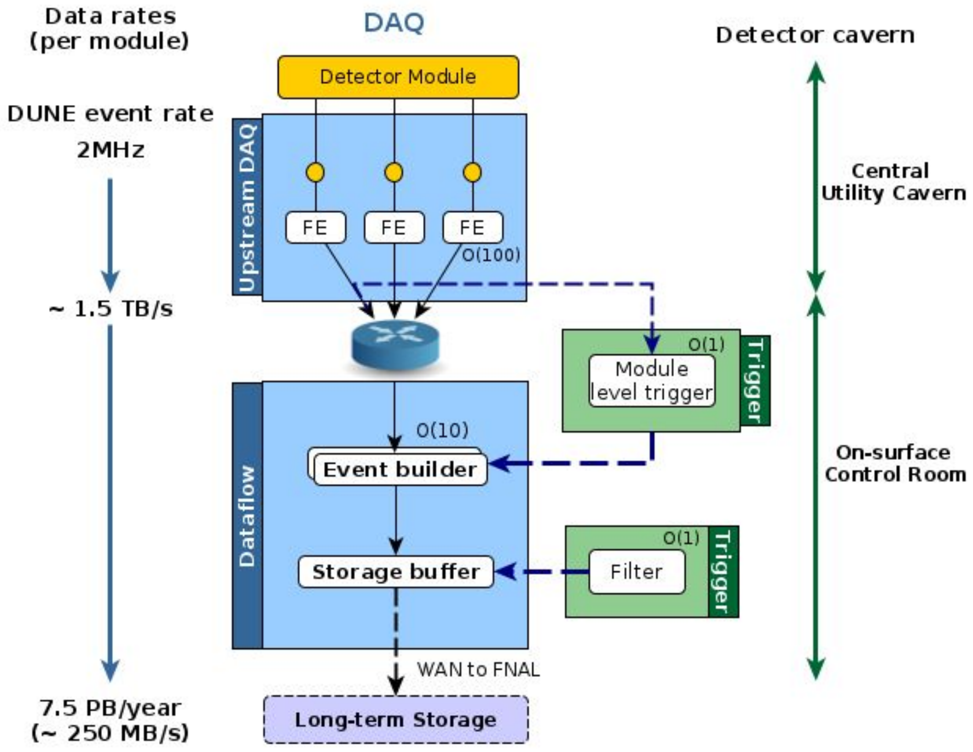
- Neutrino experiment located at Sanford Underground Research Facility in South Dakota
- Far detector located 1300 km away from source and approximately 1.48 km underground
- 4 modules of 17 kton LAr time projection chamber



DUNE experiment

DAQ system

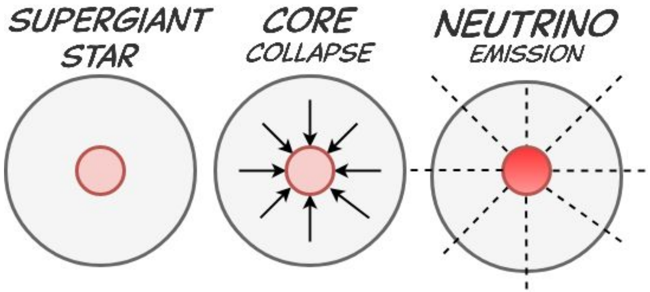
Talk on DUNE from Alessandro Thea



- 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

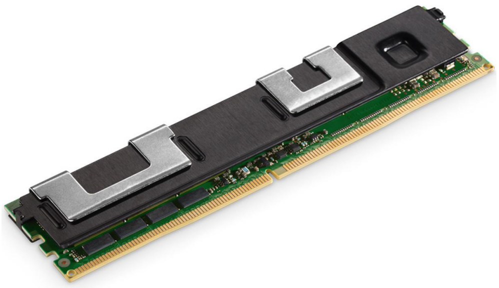
Supernova Neutrino Burst



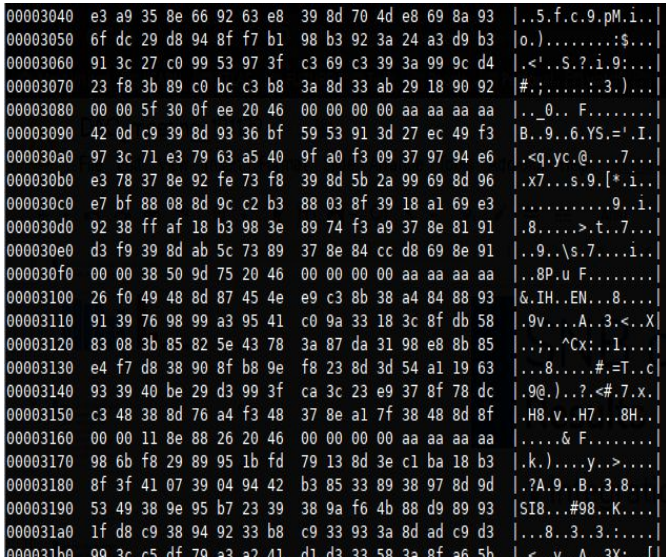
- **Supernova Neutrino Burst** (SNB) detection
  - One of the physics goals of DUNE
  - Detection of **rare**, **low energy** and **distributed** signatures
- Data taking of SNB events is **complex**:
  - Long trigger latency
  - Physics event distributed over time
  - Critical data: avoid any potential 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
- **Successful** prototype capable of buffering data from the readout system
  - Temporal buffer of 10 seconds
  - Store for over 100 seconds
  - Sustained a maximum throughput of ~ 7 GB/s
- From benchmark results: the bandwidth of NVM is approximately 10 GB/s
- **Consequence**: optimize further to exploit the full potential of the devices (**work still in progress!**)



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

UNIVERSITY OF LIVERPOOL

CERN

Thank you ! Questions ?

adam.abed.abud@cern.ch