CERN openlab ICE-DIP project: Data transfer for manycore processors

Aram Santogidis

- 23/09/2016

ICE-DIP is a European Industrial Doctorate project funded by the European Community's 7th Frameworkprogramme Marie Curie Actions under grant PITN-GA-2012-316596

Background image: Shutterstock

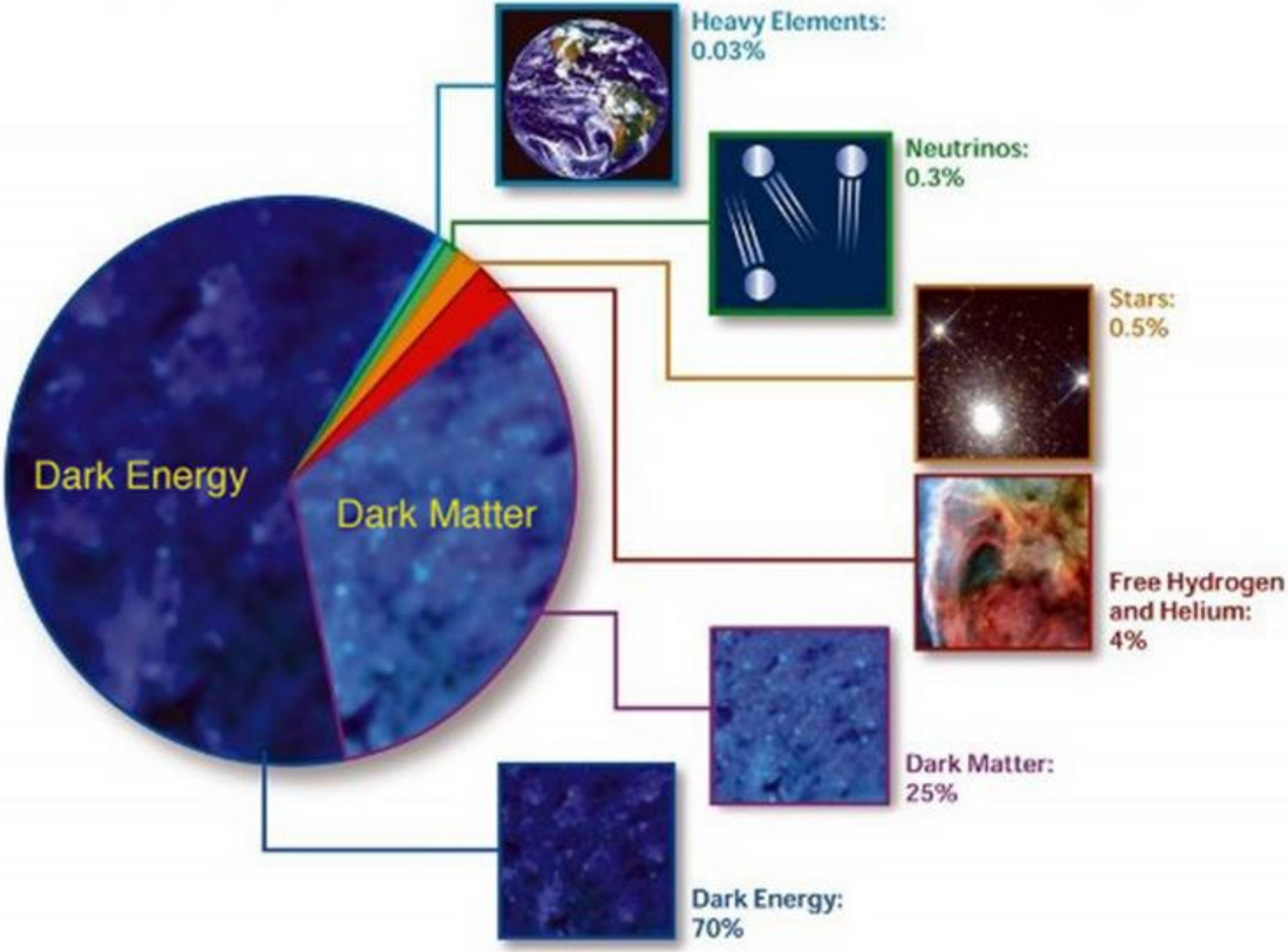
Looking at the sky…

Background image: Shutterstock

… more like this

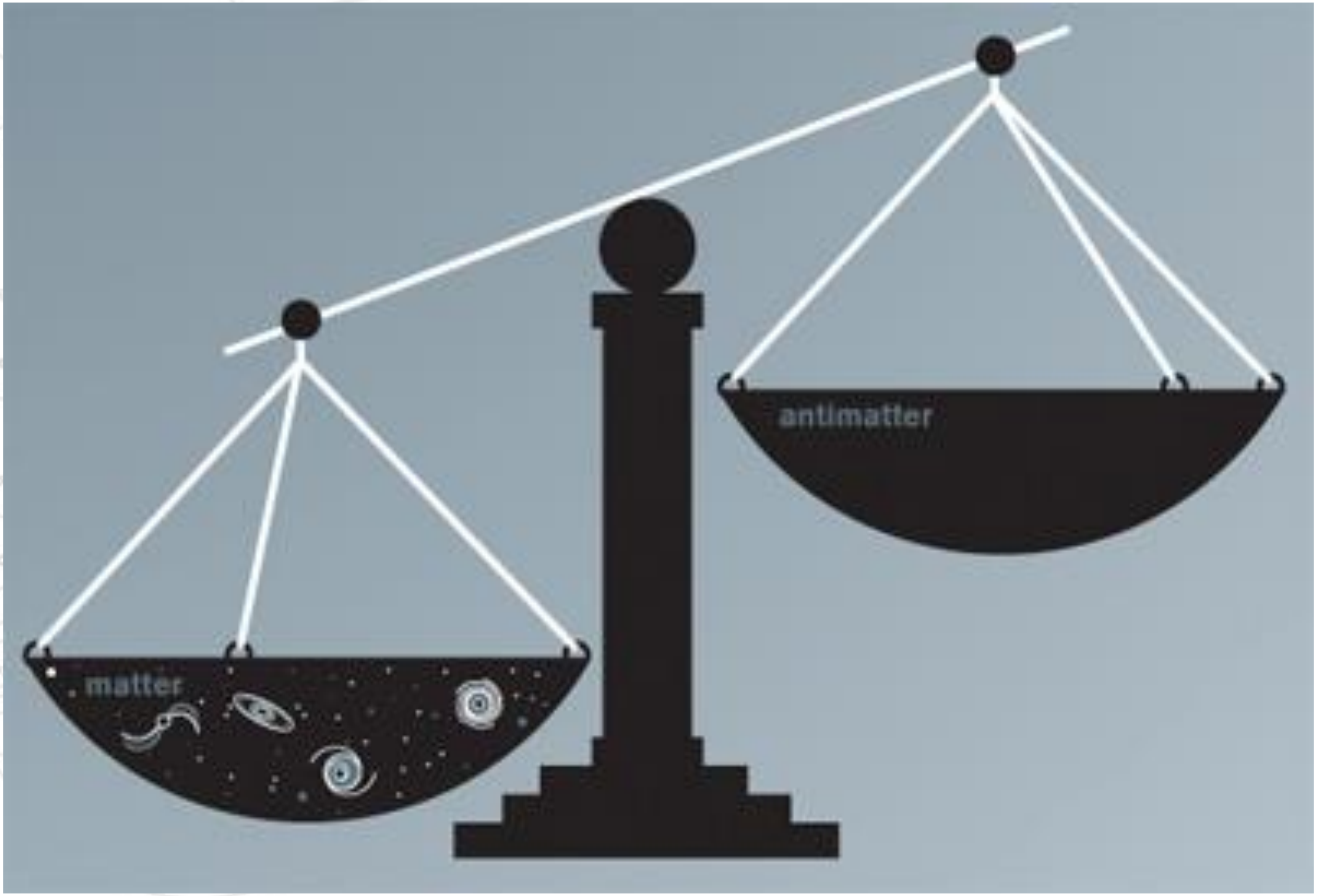
Background image: Shutterstock

Big questions about the Universe



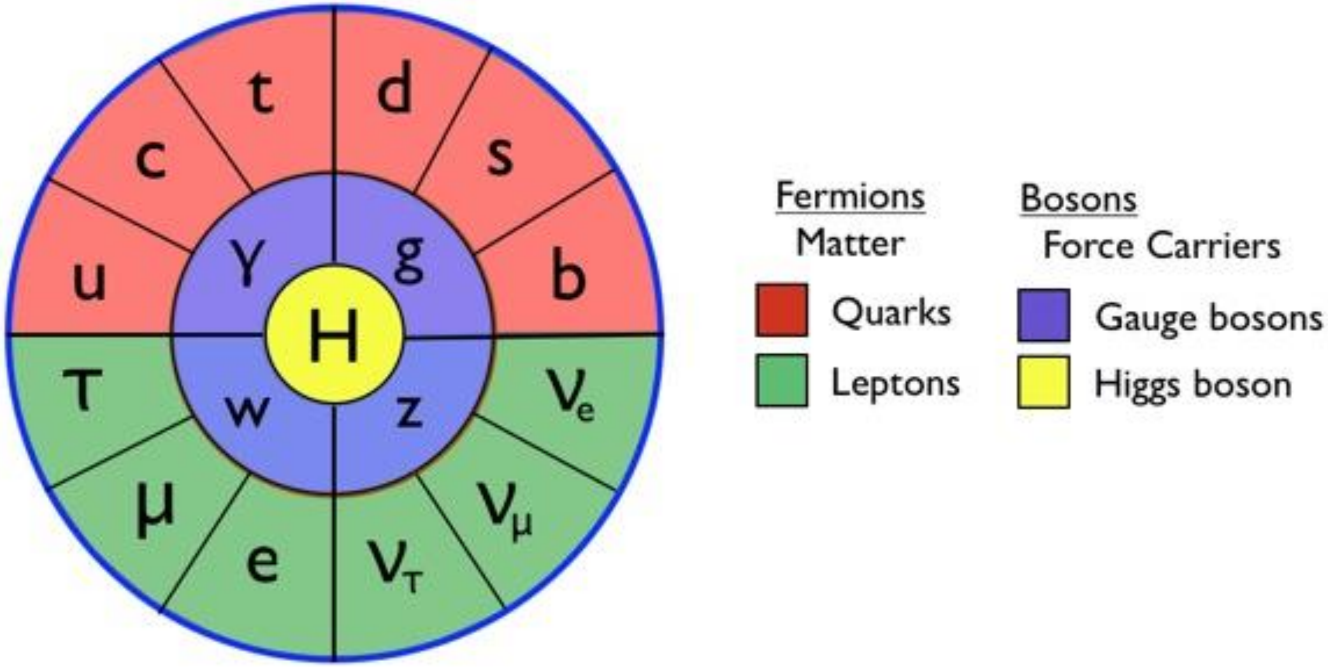
Background image: Shutterstock

Charge-Parity violation



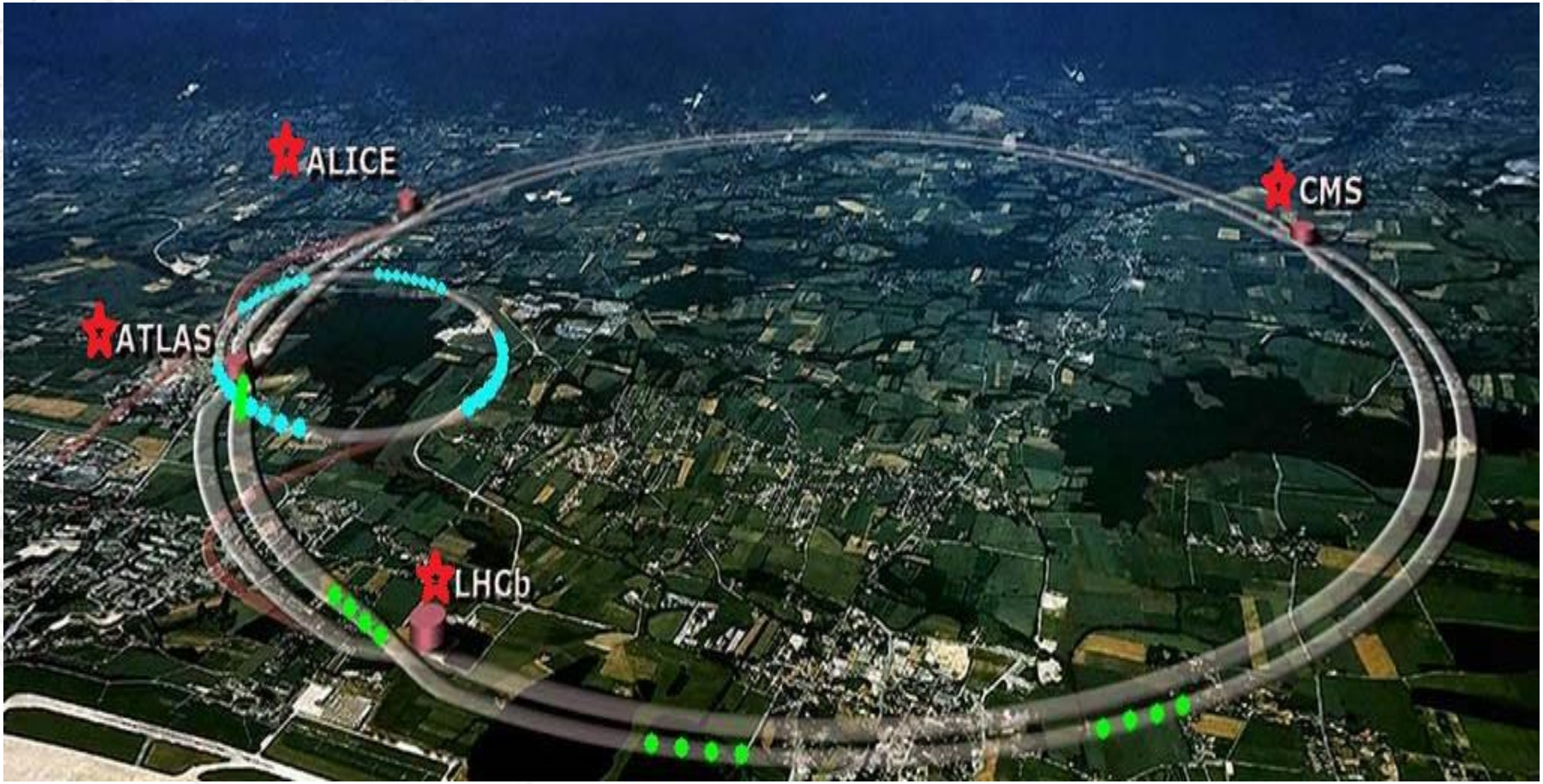
Background image: Shutterstock

Higgs bosson



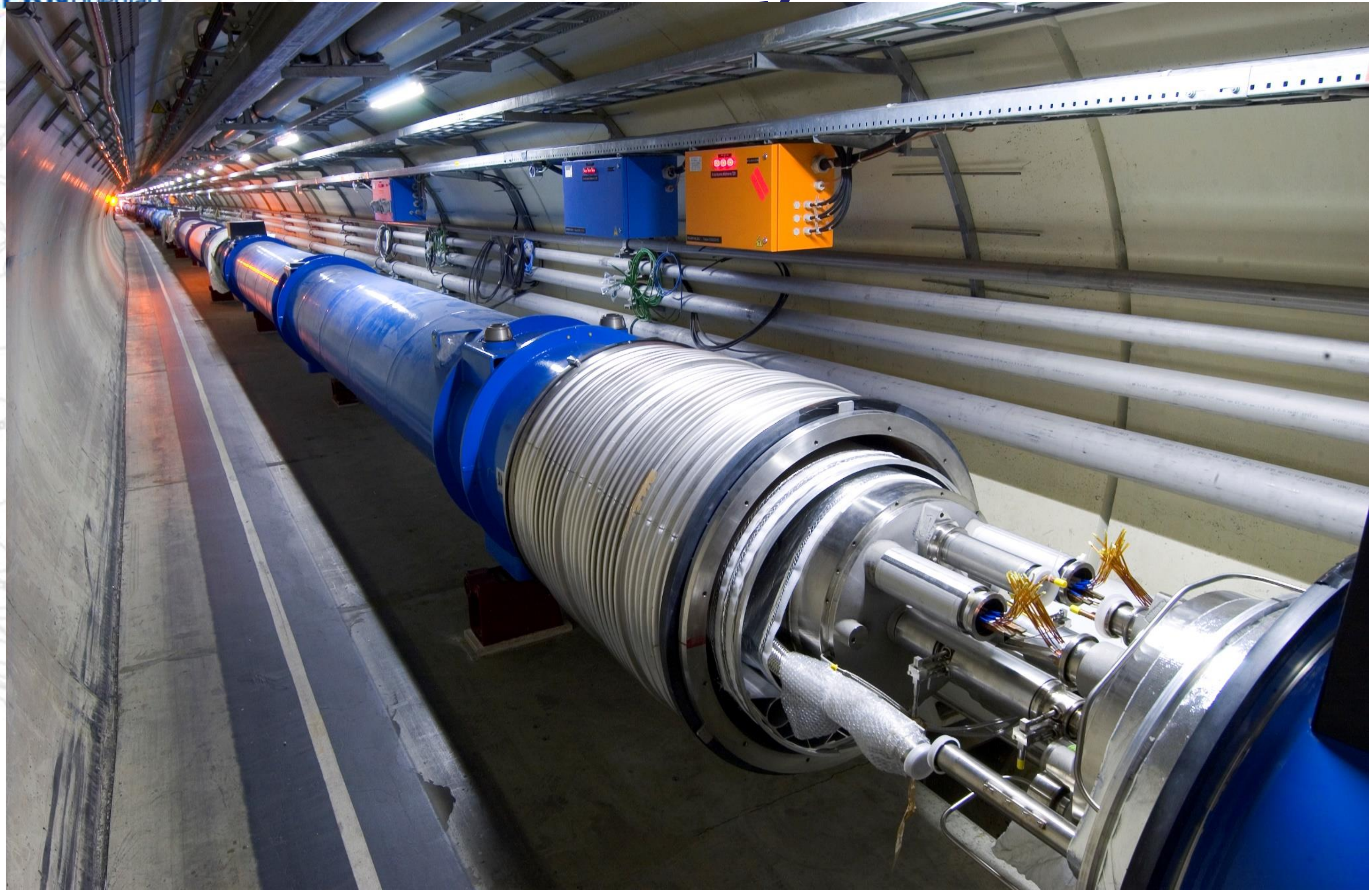
Particles of the Standard Model

Physics Experiments at CERN



Background image: Shutterstock

The Large Hadron Collider

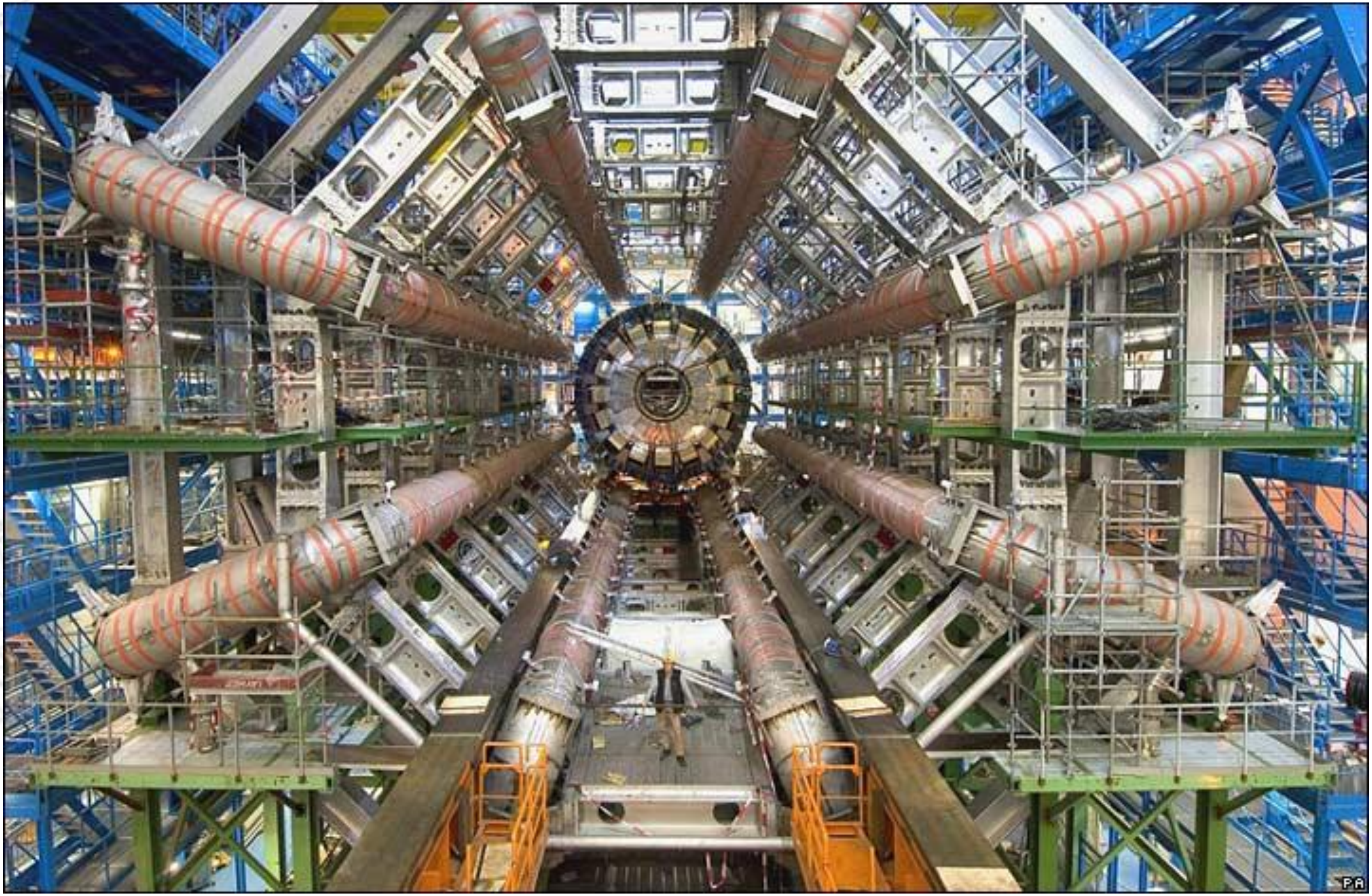


Background image: Shutterstock

Engineers maintaining LHC

Background image: Shutterstock

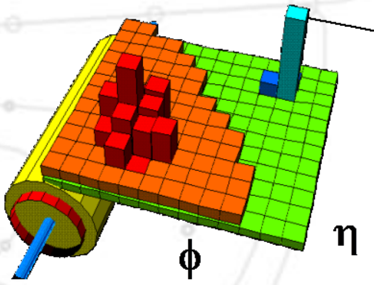
The ATLAS experiment detector

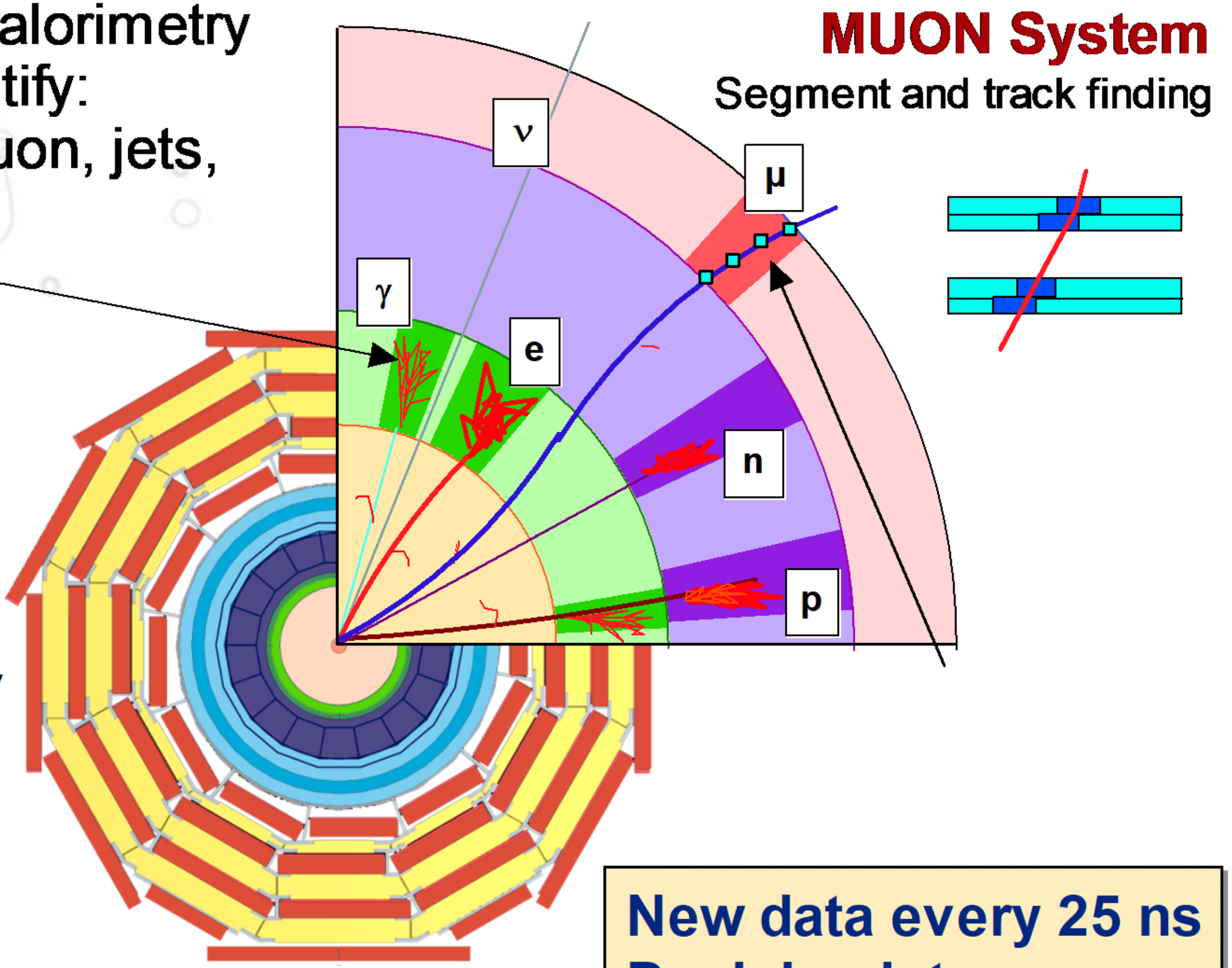


Background image: Shutterstock

Particle trajectories in the detector

Use prompt data (calorimetry and muons) to identify:
High p_t electron, muon, jets, missing E_T





MUON System

Segment and track finding

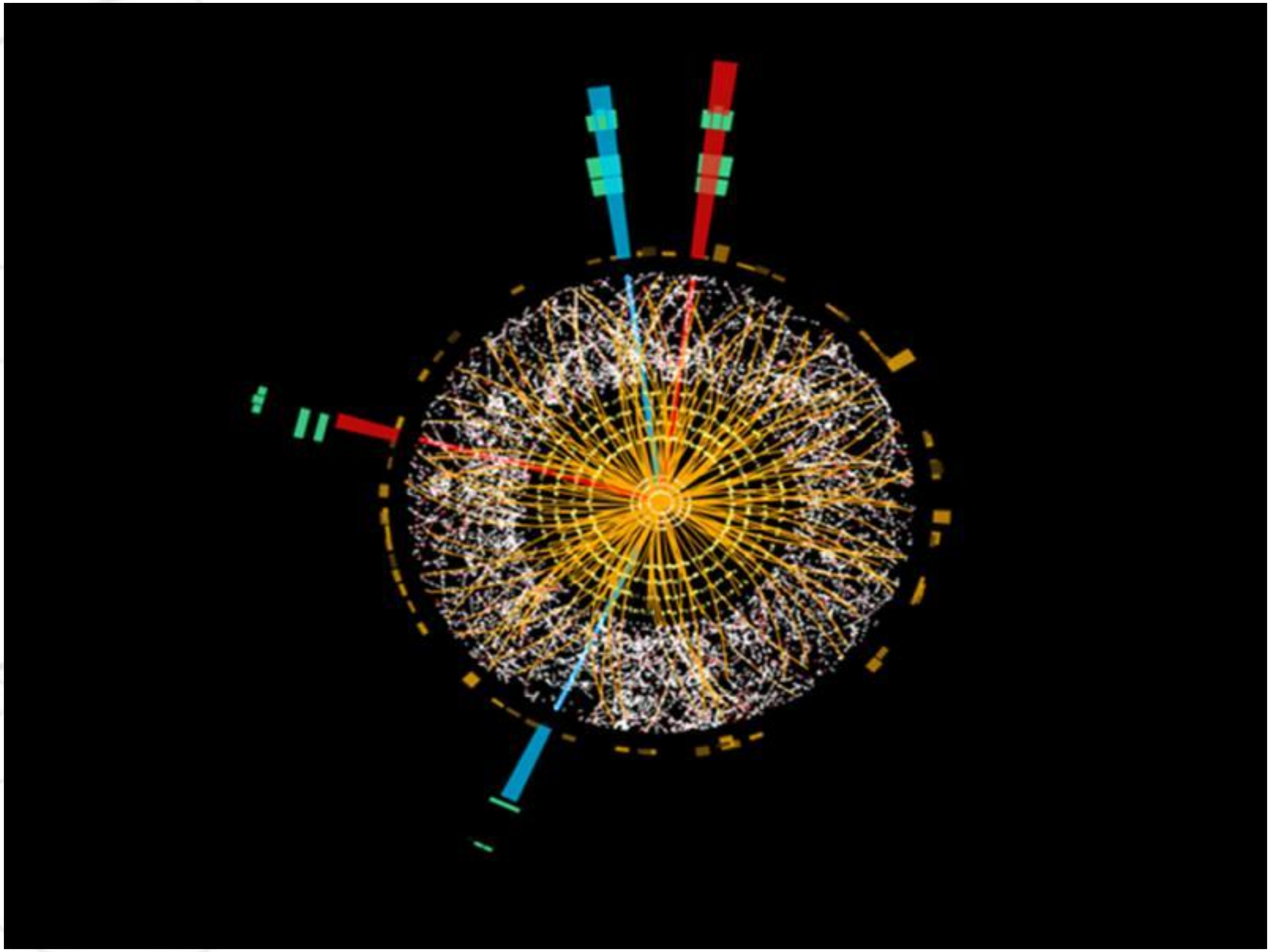
CALORIMETERs

Cluster finding and energy deposition evaluation

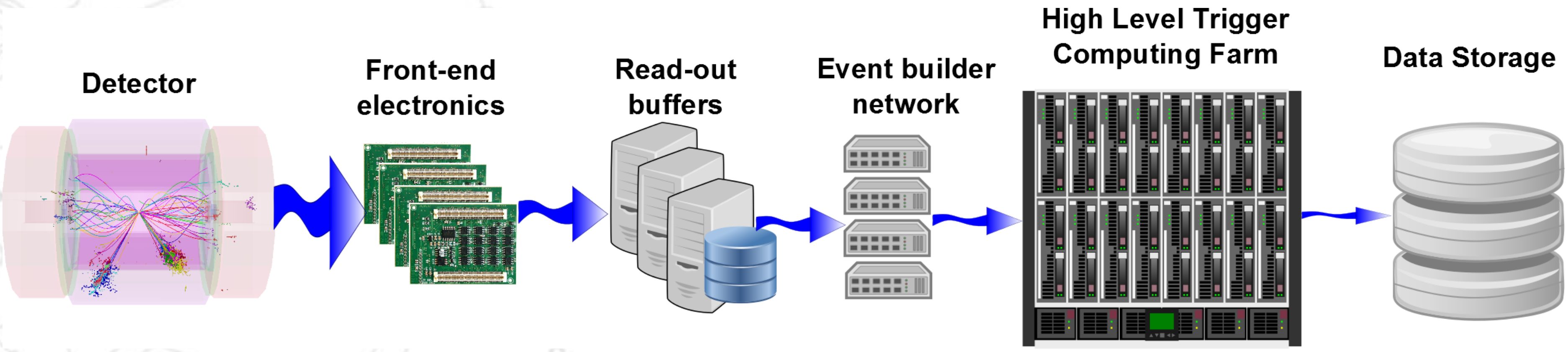
New data every 25 ns
Decision latency ~ μs

Background image: Shutterstock

A large number of particle trajectories



Online processing system

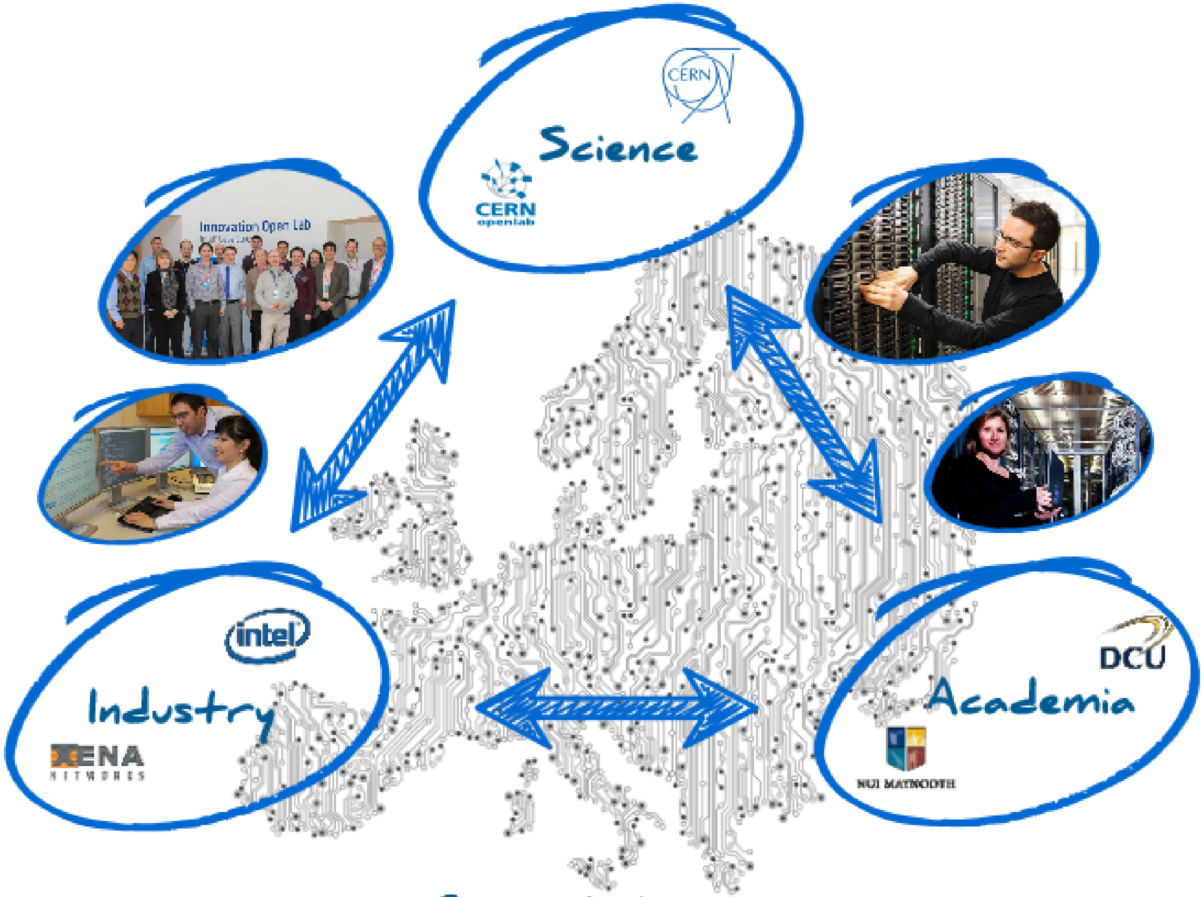


1 PB/s to 600 MB/s

Background image: Shutterstock

ICE-DIP 2013-2017:
The Intel-CERN European Doctorate Industrial Program

A public-private partnership to research solutions for next generation data acquisition networks, offering research training to five Early Stage Researchers in ICT



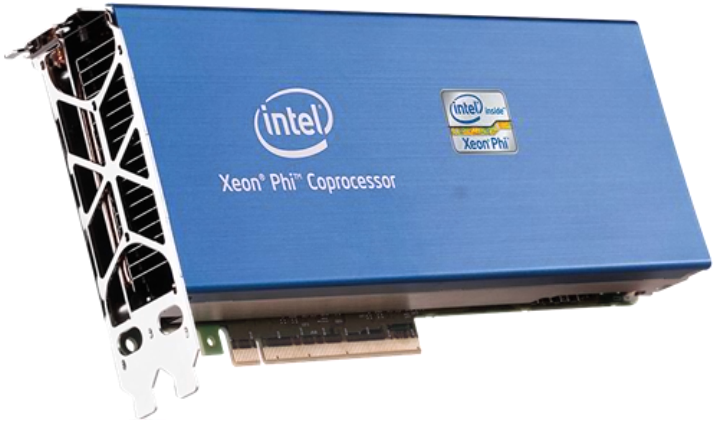
Research topics:

- Silicon photonics systems
- Next generation data acquisition networks
- High speed configurable logic
- Computing solutions for high performance data filtering

Background image: Shutterstock

Intel Xeon Phi Coprocessor

- **Up to 61 Cores**
- **PCIe card**
- **Different computation modes**
  - Offloading
  - Symmetric



Background image: Shutterstock

My focus in the online processing pipeline

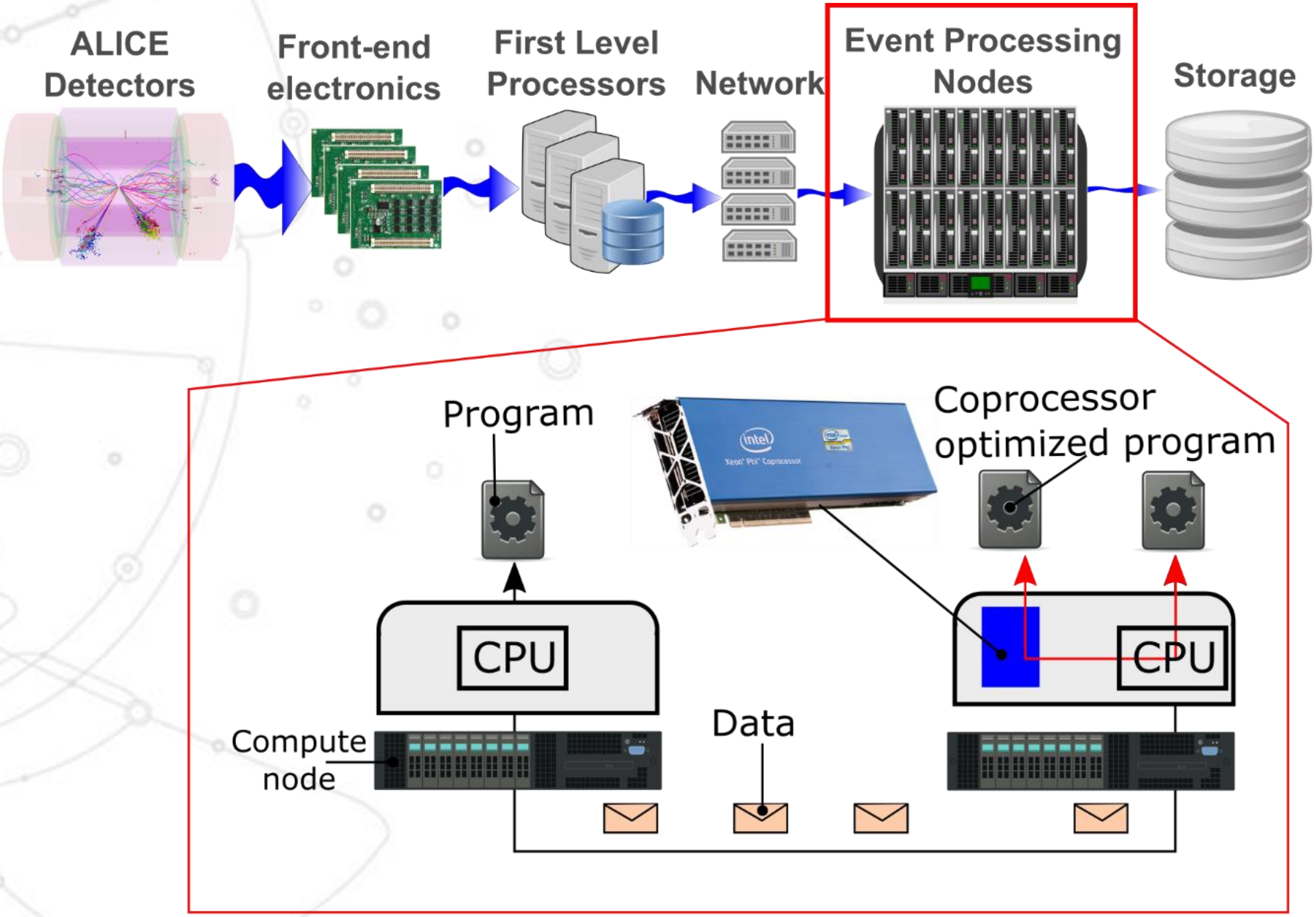
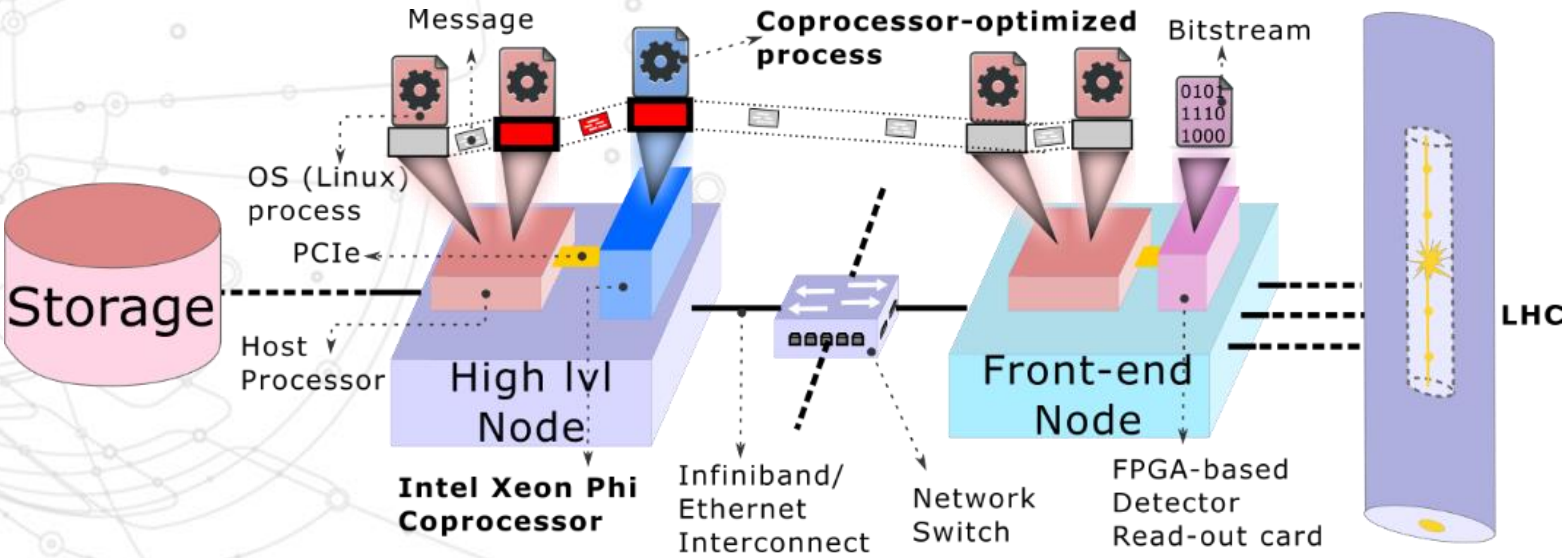


Figure 1. An overview of the ALICE data acquisition process with emphasis on the Event Processing Nodes.

Coprocessor-Host communication in the online processing system


Message
Coprocessor-optimized process
Bitstream
0101
1110
1000
OS (Linux) process
PCIe
Storage
Host Processor
High lvl Node
Intel Xeon Phi Coprocessor
Infiniband/ Ethernet Interconnect
Network Switch
Front-end Node
FPGA-based Detector Read-out card
LHC

Data transfer libraries for the data acquisition system of the ALICE experiment

ØMQ

nanomsg

- High level messaging patterns (e.g. pub-sub)
- Used to create distributed systems
- Provides good performance

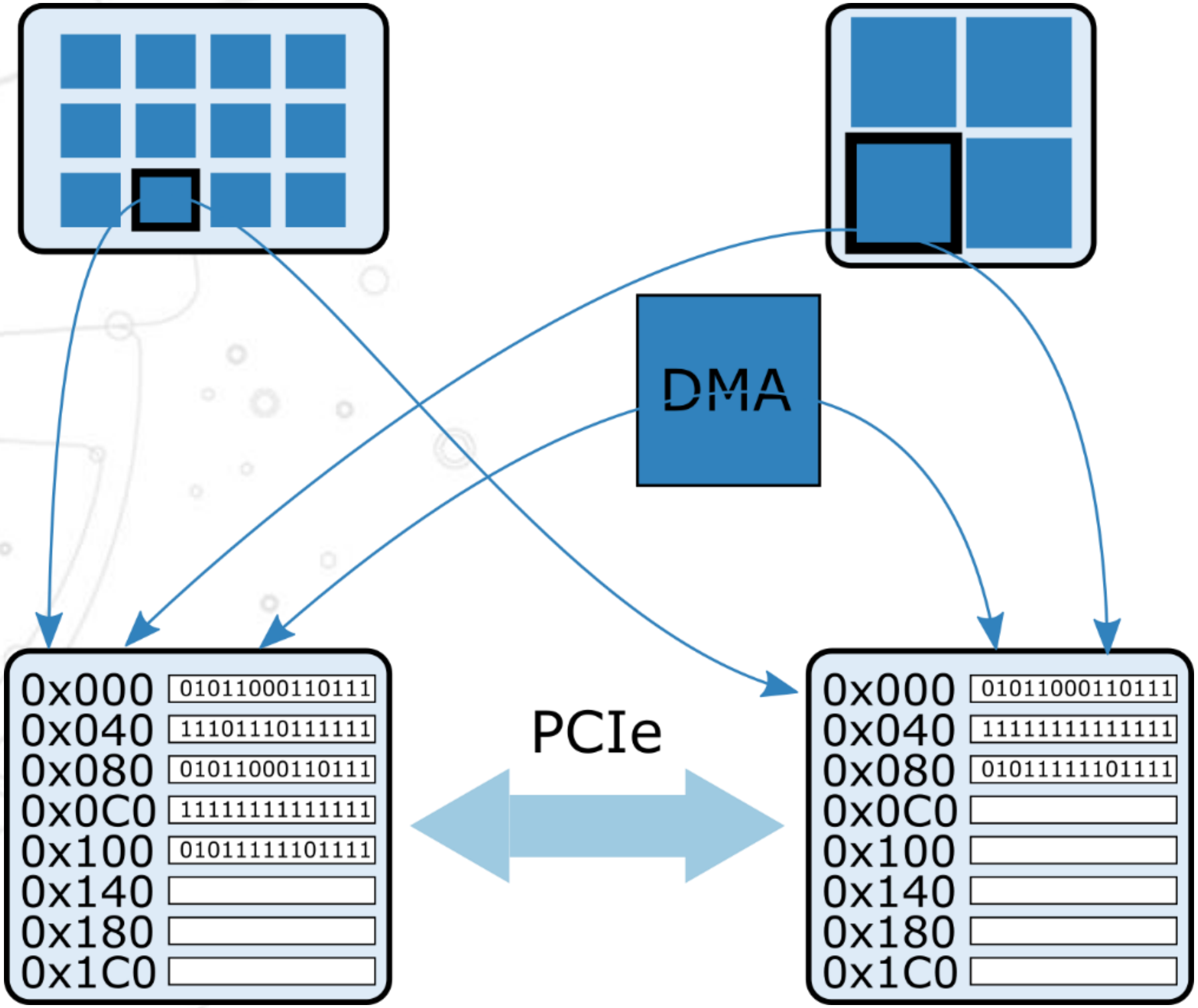
Background image: Shutterstock

Symmetric Communications Interface (SCIF) For Intel® Xeon Phi

- Communication over PCIe with minimum overhead
- POSIX-like interface (listen, connect, send, receive)
- RDMA capability (registering memory and read/write to remote address space)
- memory mapped IO along the lines of POSIX mmap

Background image: Shutterstock

SCIF RDMA

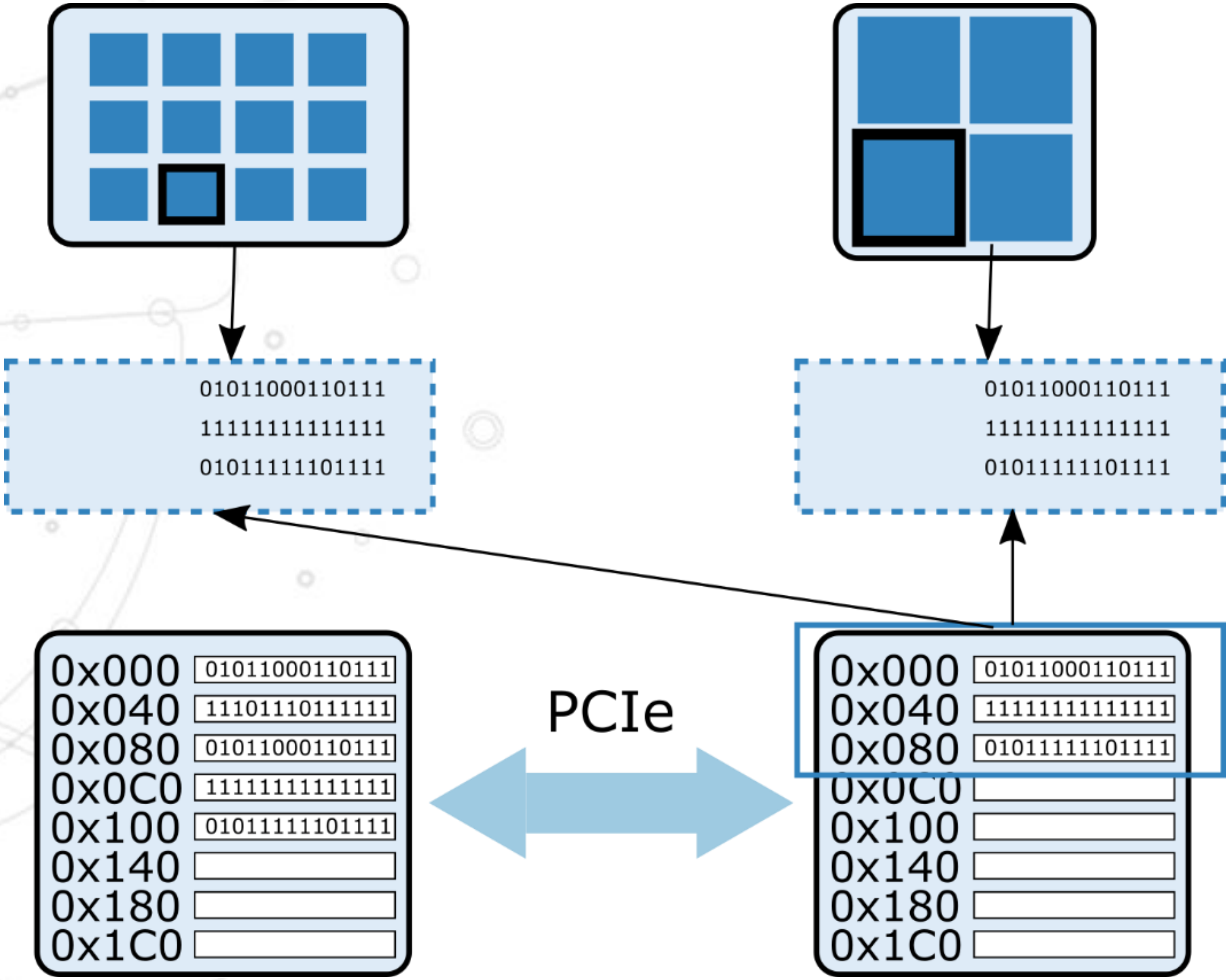


0x000	01011000110111
0x040	11101110111111
0x080	01011000110111
0x0C0	11111111111111
0x100	01011111101111
0x140	
0x180	
0x1C0	

0x000	01011000110111
0x040	11111111111111
0x080	01011111101111
0x0C0	
0x100	
0x140	
0x180	
0x1C0	

Background image: Shutterstock

SCIF mmap


01011000110111
11111111111111
01011111101111
01011000110111
11111111111111
01011111101111
0x000 01011000110111
0x040 11101110111111
0x080 01011000110111
0x0C0 11111111111111
0x100 01011111101111
0x140
0x180
0x1C0
PCIe
0x000 01011000110111
0x040 11111111111111
0x080 01011111101111
0x0C0
0x100
0x140
0x180
0x1C0

An opportunity for improvement

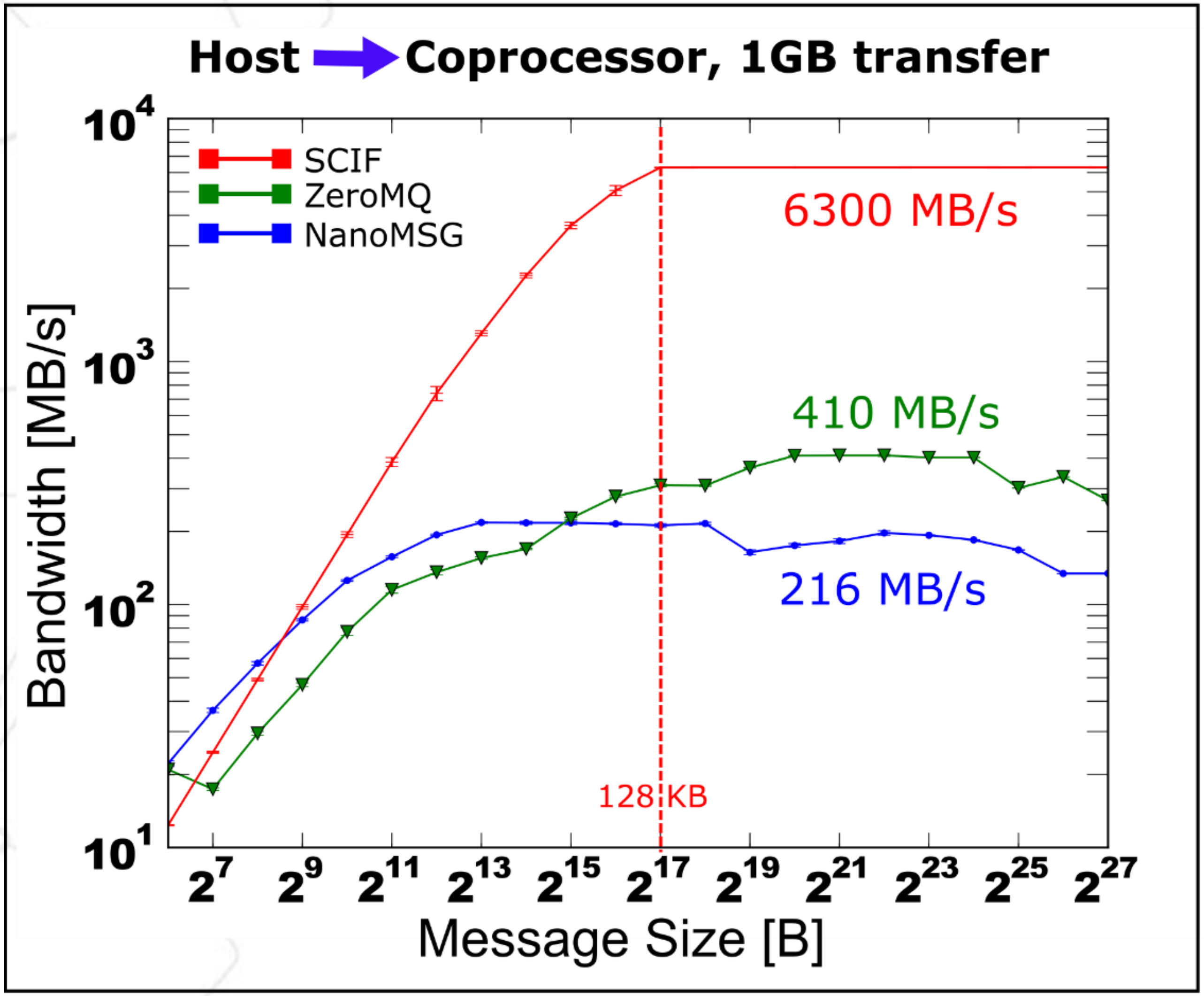


***Figure 2** . This plot corresponds toa performance test of transfer-ring1GB payload in chunks from 4KB to 128MB[1].*

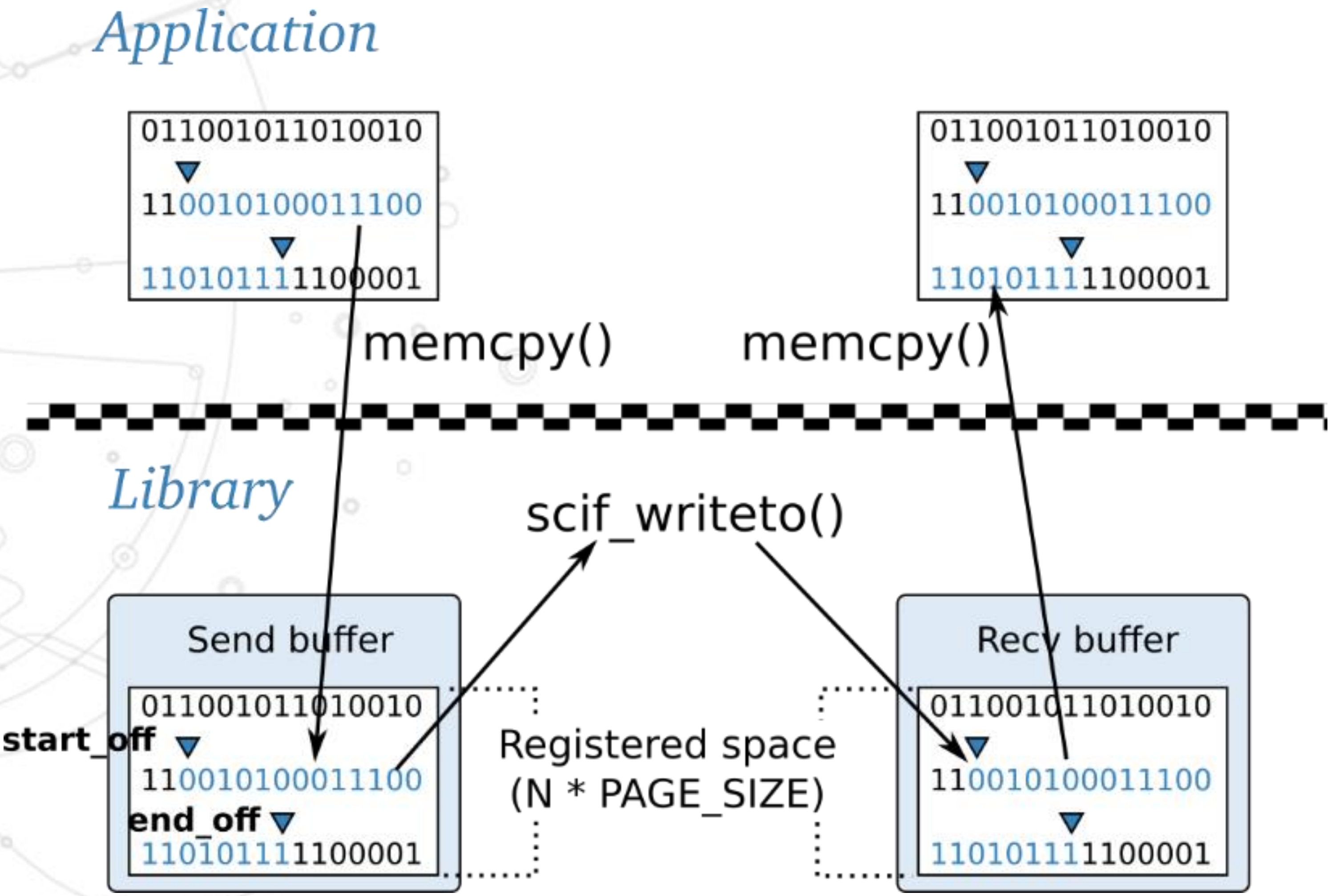
Background image: Shutterstock

The features of the new transport mechanism over SCIF

- Streaming semantics (along the lines of TCP)
- Cacheline-aligned RDMA transfers only
- Lock-free one-sided communication
- Maximize data transfer throughput (bandwidth)

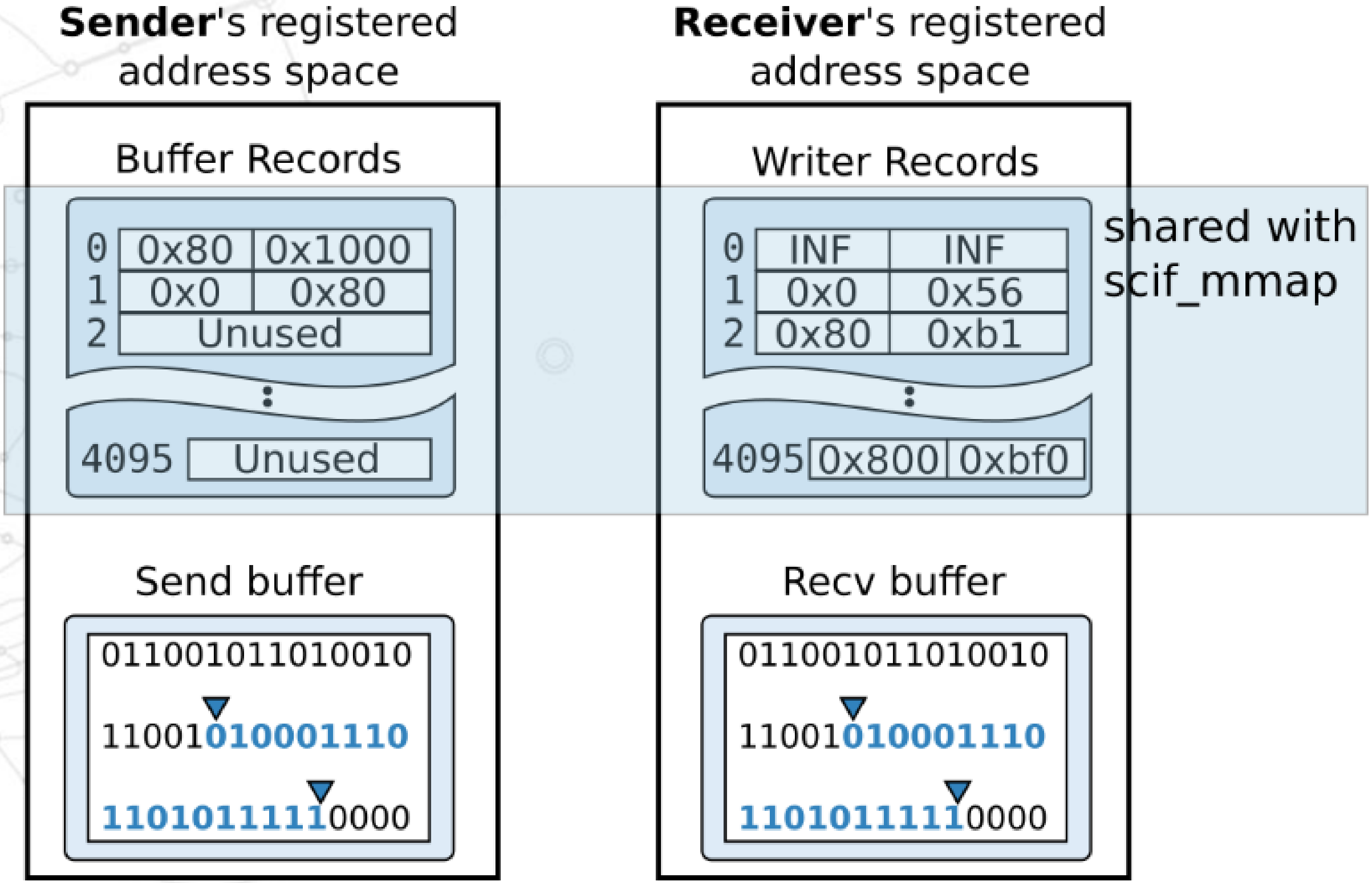
Background image: Shutterstock

Trans4SCIF synchronization algorithm



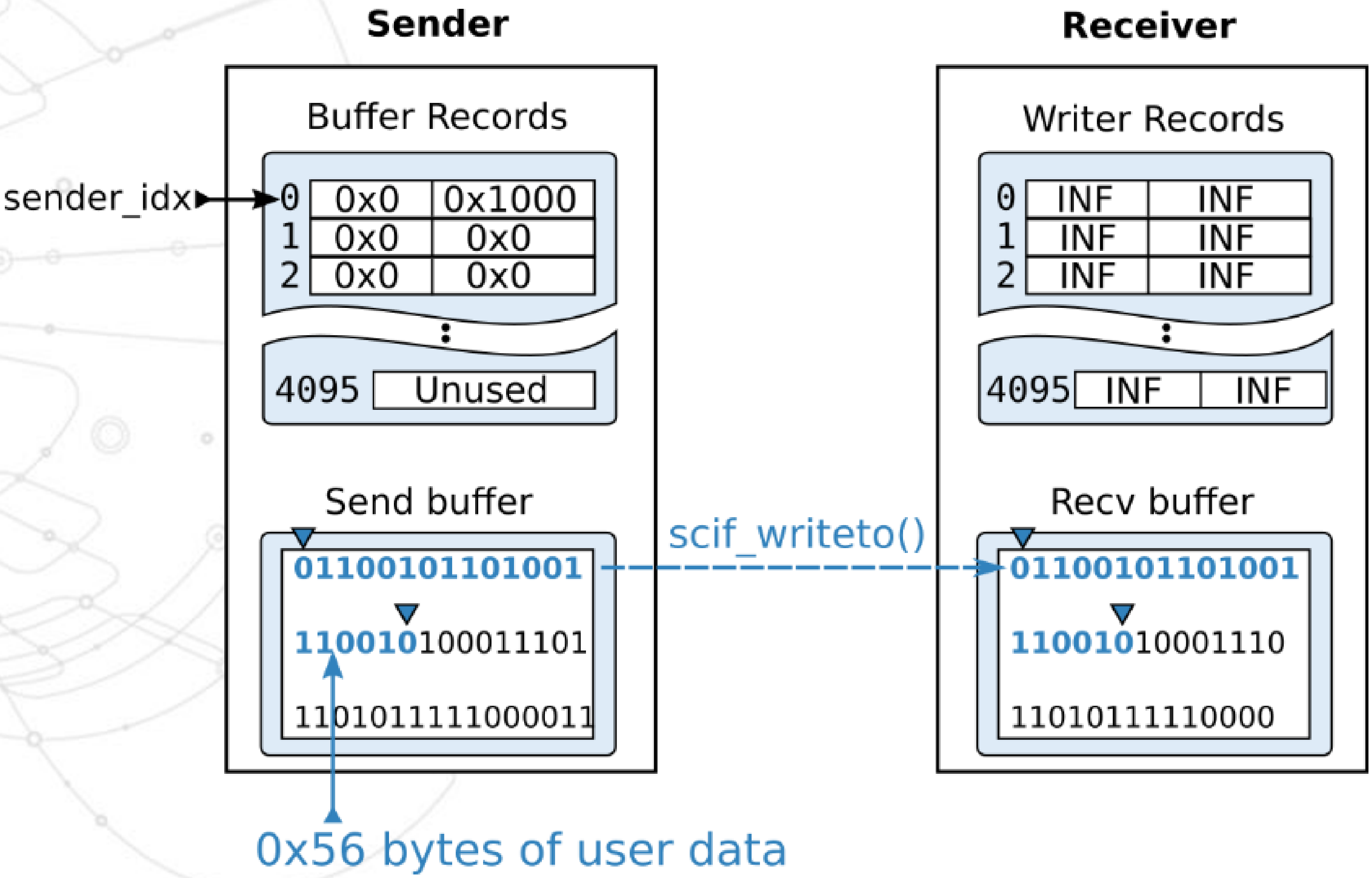
Background image: Shutterstock

Trans4SCIF synchronization algorithm



Background image: Shutterstock

Trans4SCIF synchronization algorithm



Sender

Buffer Records

sender_idx

0	0x0	0x1000
1	0x0	0x0
2	0x0	0x0
⋮		
4095	Unused	

Send buffer

01100101101001

110010100011101

11010111111000011

0x56 bytes of user data

scif_writeto()

Receiver

Writer Records

0	INF	INF
1	INF	INF
2	INF	INF
⋮		
4095	INF	INF

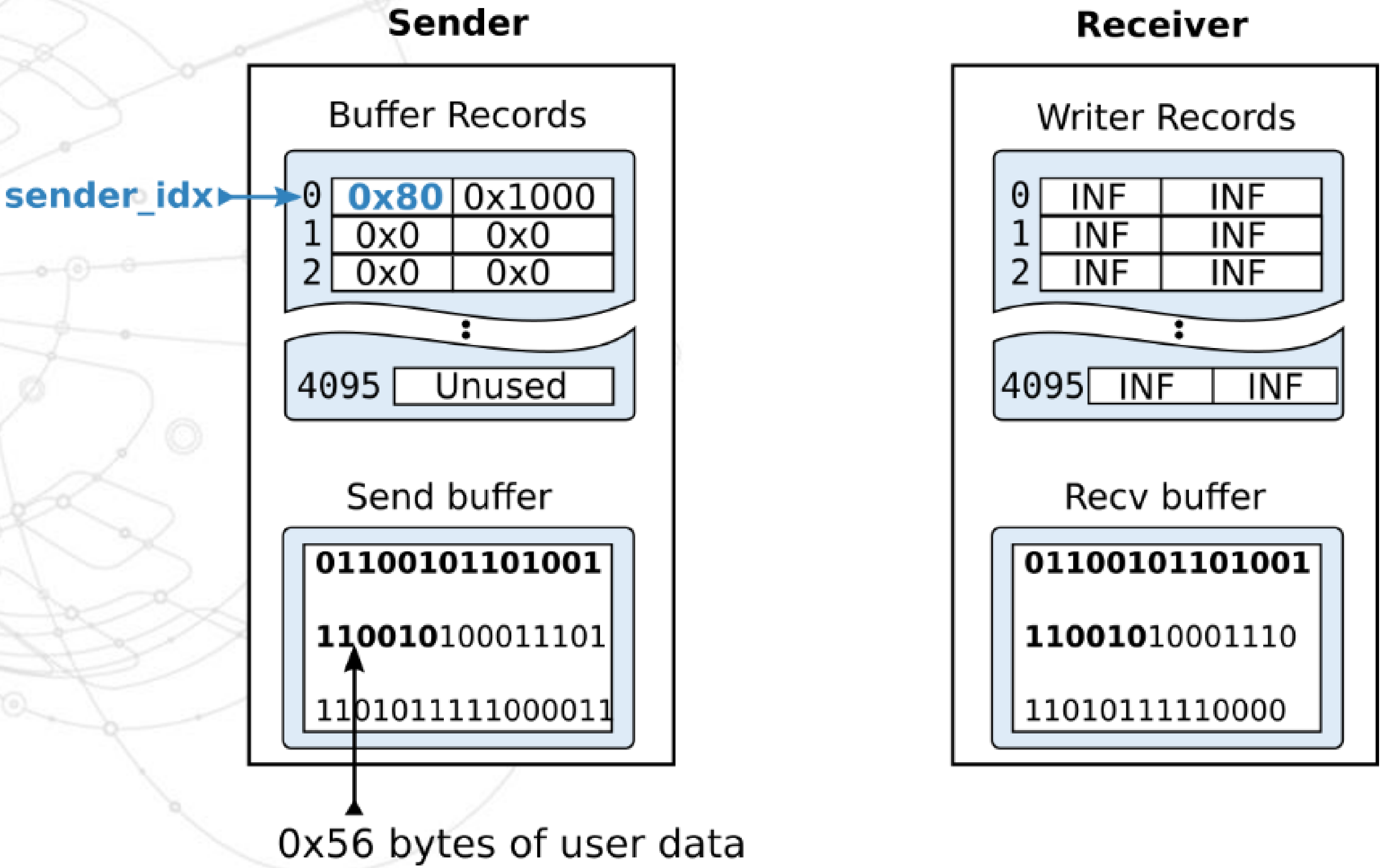
Recv buffer

01100101101001

11001010001110

11010111110000

Trans4SCIF synchronization algorithm



Sender

Buffer Records

sender_idx

0	0x80	0x1000
1	0x0	0x0
2	0x0	0x0
⋮		
4095	Unused	

Send buffer

01100101101001
110010100011101
110101111100011

0x56 bytes of user data

Receiver

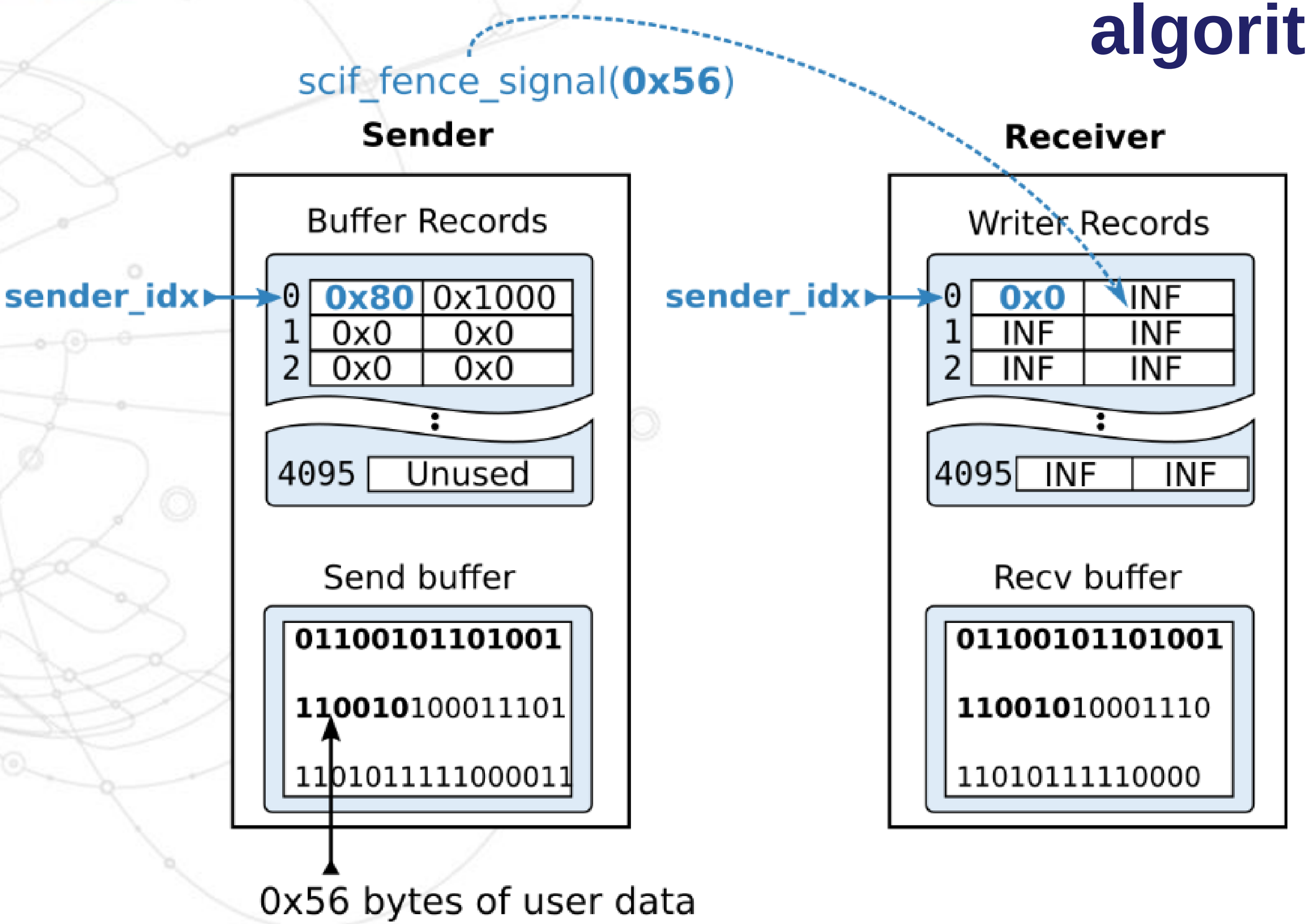
Writer Records

0	INF	INF
1	INF	INF
2	INF	INF
⋮		
4095	INF	INF

Recv buffer

01100101101001
11001010001110
11010111110000

Trans4SCIF synchronization algorithm



Buffer Records

0	**0x80**	0x1000
1	0x0	0x0
2	0x0	0x0
⋮		
4095	Unused	

Writer Records

0	**0x0**	INF
1	INF	INF
2	INF	INF
⋮		
4095	INF	INF

Send buffer

01100101101001

110010100011101

11010111111000011

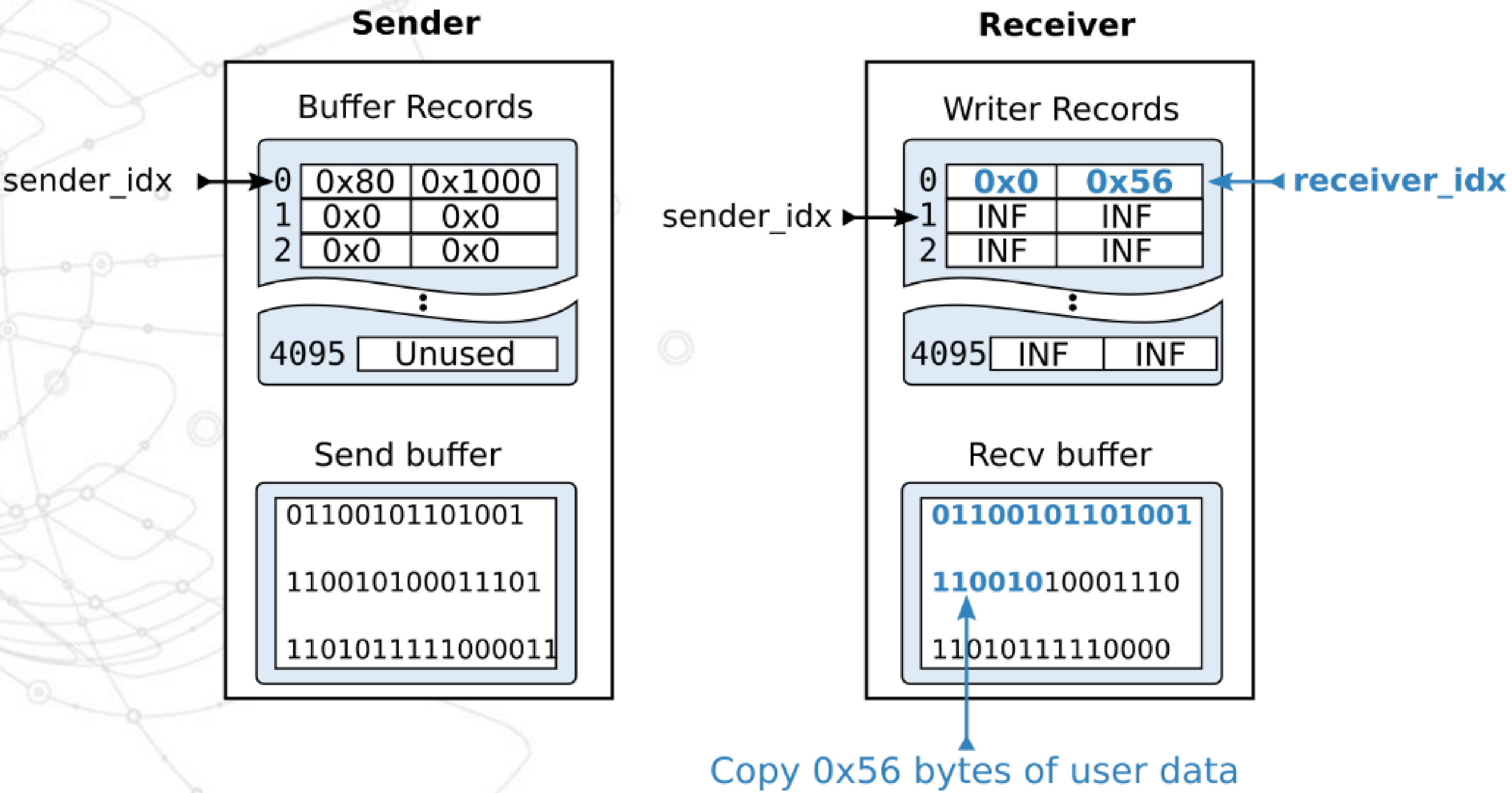
Recv buffer

01100101101001

11001010001110

11010111110000

Trans4SCIF synchronization algorithm



Sender

sender_idx → 0

Buffer Records

0	0x80	0x1000
1	0x0	0x0
2	0x0	0x0
⋮		
4095	Unused	

Send buffer

```
01100101101001

110010100011101

1101011111000011
```

Receiver

sender_idx → 1

receiver_idx → 0

Writer Records

0	**0x0**	**0x56**
1	INF	INF
2	INF	INF
⋮		
4095	INF	INF

Recv buffer

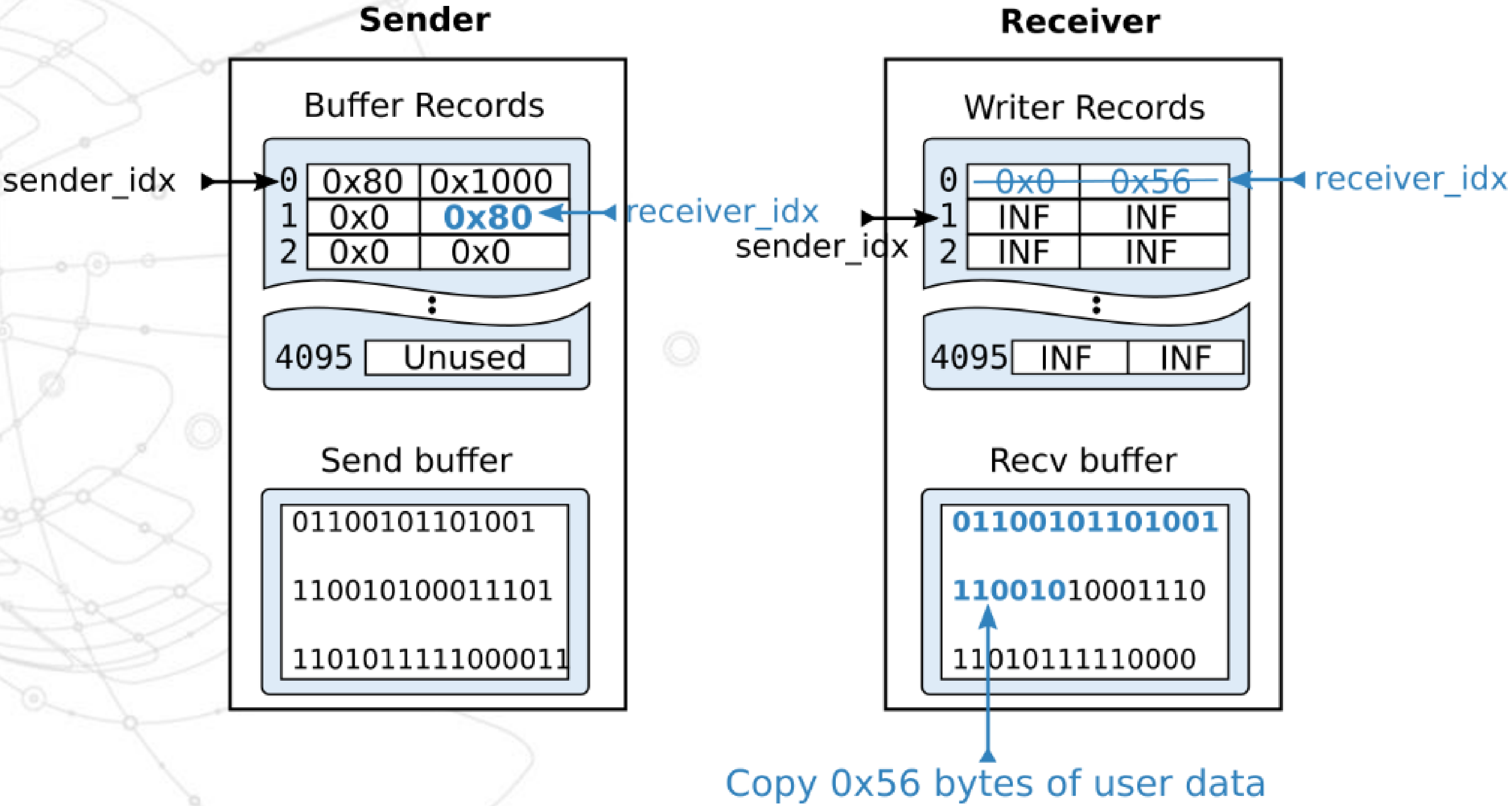
```
01100101101001

11001010001110

11010111110000
```

Copy 0x56 bytes of user data

Trans4SCIF synchronization algorithm



Sender

sender_idx

Buffer Records

0	0x80	0x1000
1	0x0	**0x80**
2	0x0	0x0
⋮		
4095	Unused	

receiver_idx

Send buffer

```
01100101101001
110010100011101
110101111100011
```

Receiver

sender_idx

Writer Records

0	~~0x0~~	~~0x56~~
1	INF	INF
2	INF	INF
⋮		
4095	INF	INF

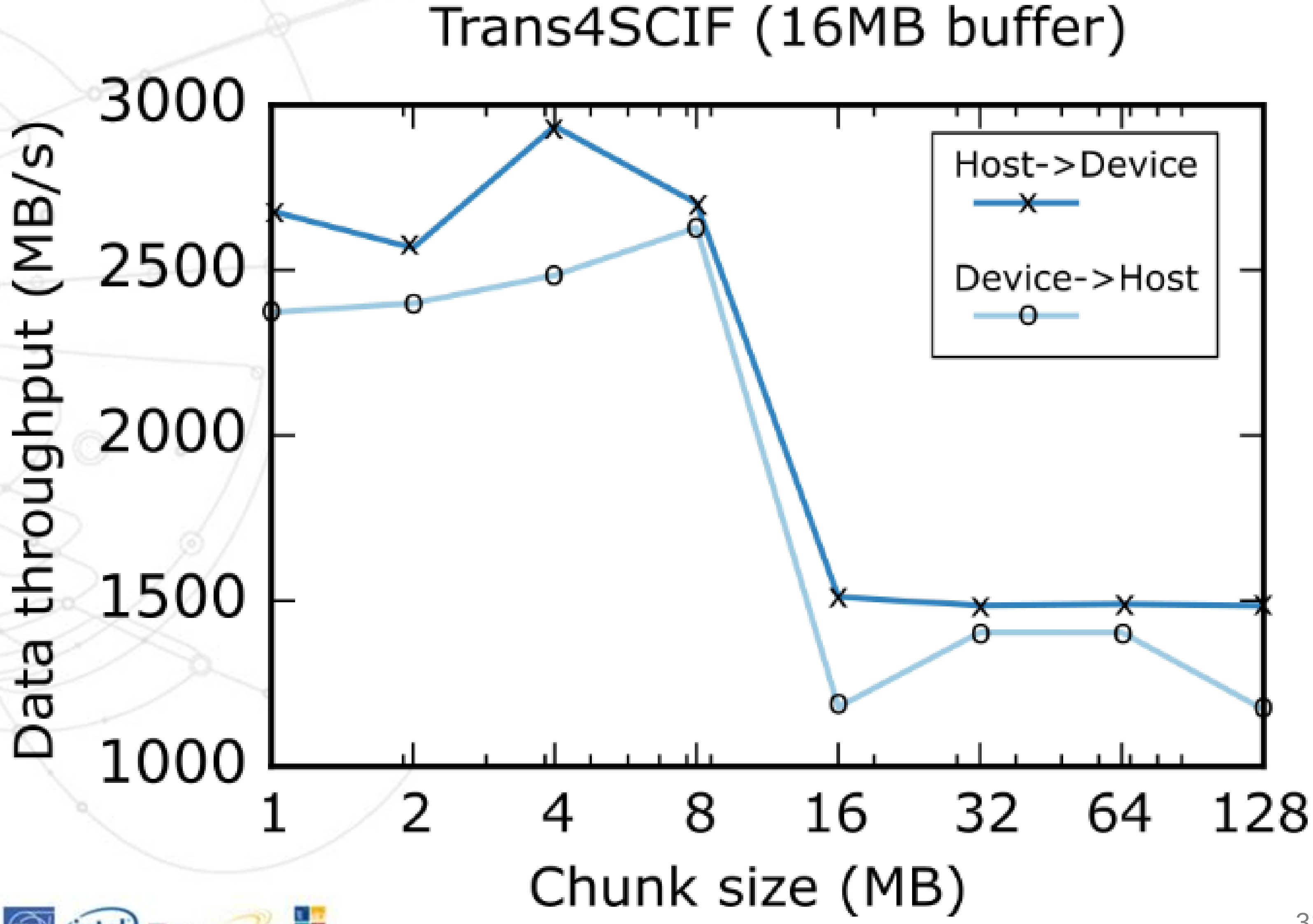
receiver_idx

Recv buffer

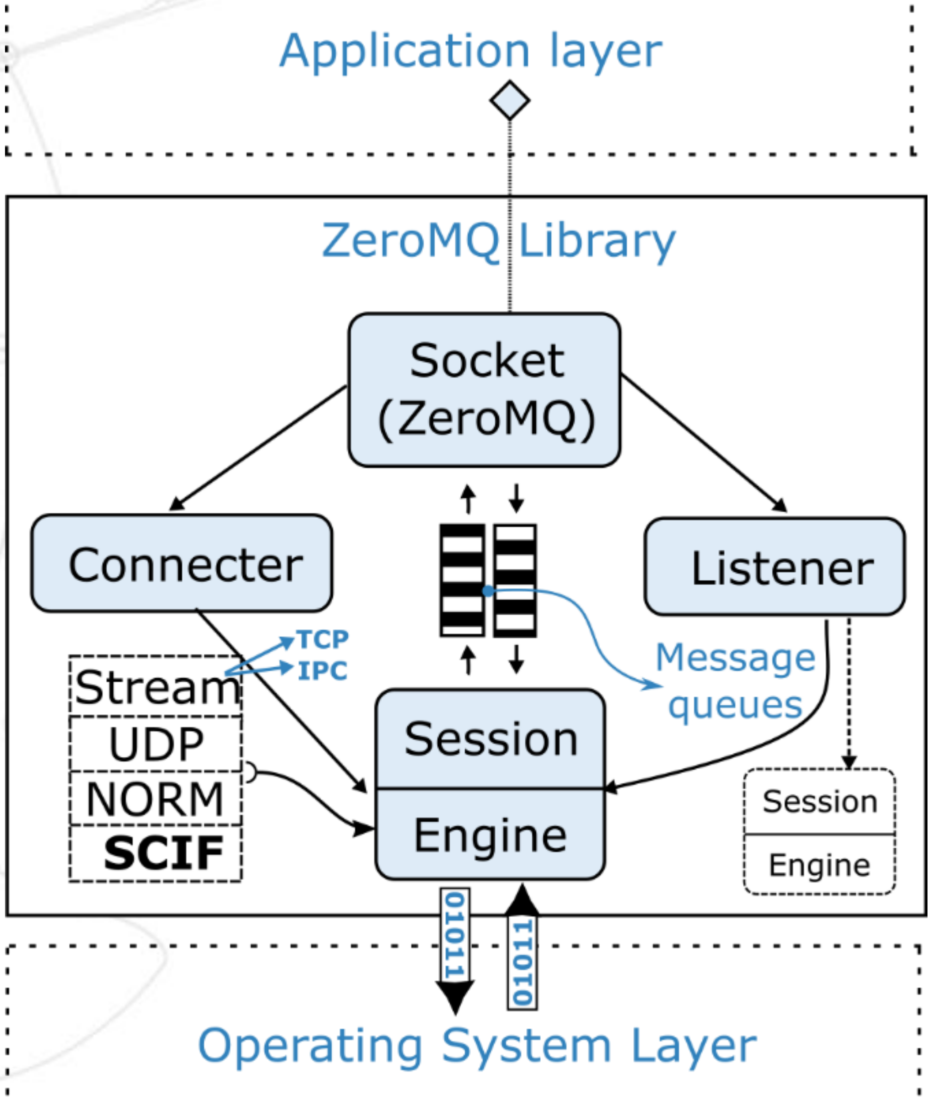
```
01100101101001
11001010001110
11010111110000
```

Copy 0x56 bytes of user data

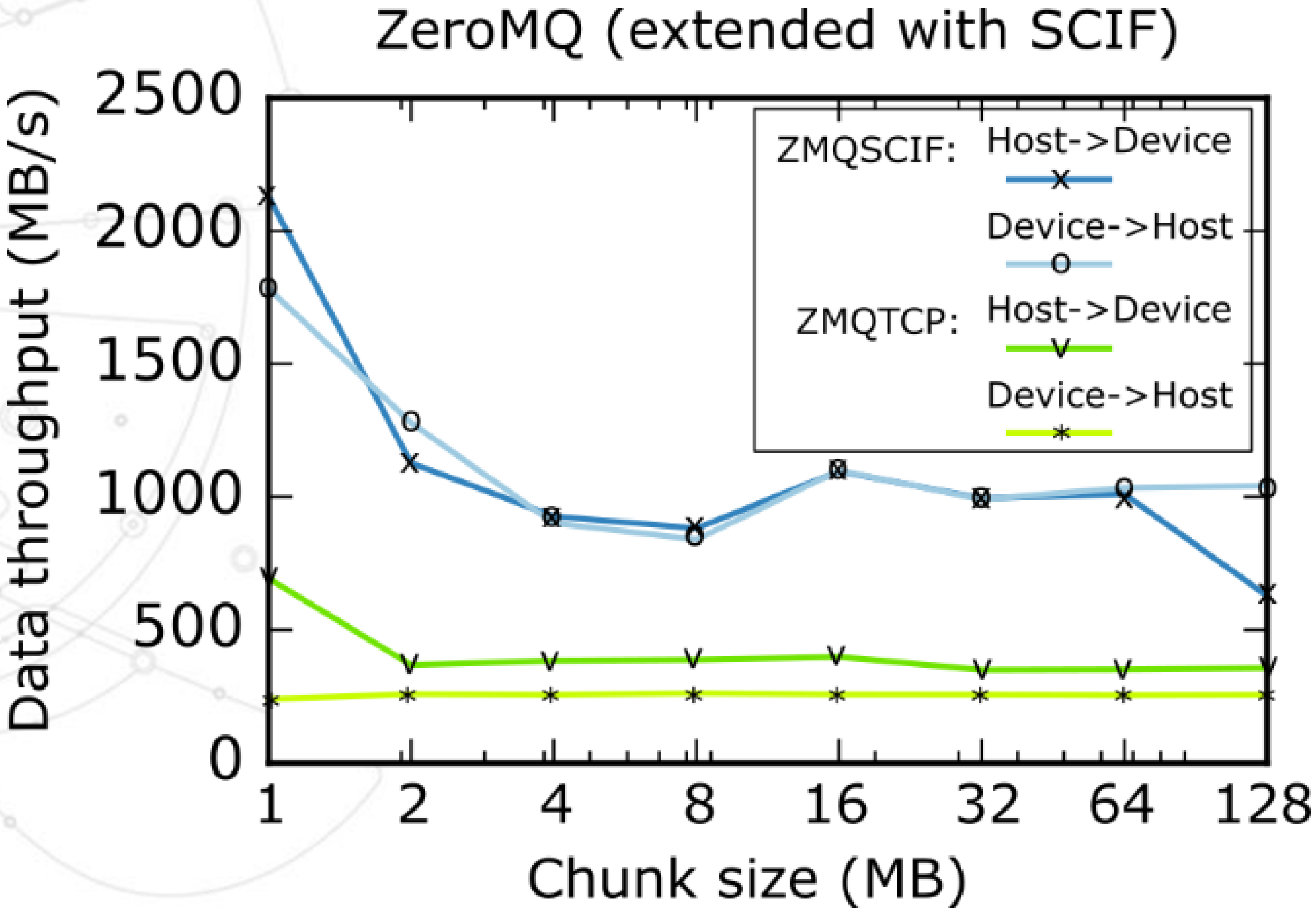
The performance of Trans4SCIF



The ZeroMQ extension with SCIF



The performance of ZeroMQ extend with SCIF via Trans4SCIF



The Trans4SCIF library the Intel Xeon Phi Coprocessor

- **Easy-to-use socket-like interface**
  - Send/Recv
- **E.g. up to 3 GB/s data throughput (4x imporovementwith the cost of 32 MB of memory space**
- **ZeroMQ extension for SCIF**
- **In principle can be re-used by other RDMA based transports (e.g OmniPath)**

Background image: Shutterstock

Acknowledgements

Than you for your attention.

This research project has been supported by a Marie Curie Early European Industrial Doctorates Fellowship of the European Community's Seventh Framework Programme under contract number (PITN-GA-2012-316596-ICE-DIP).

Background image: Shutterstock