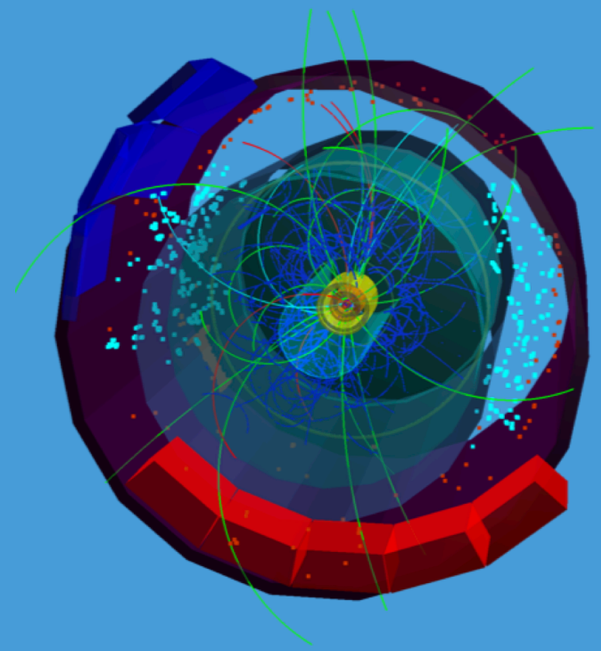
17 Apr 2024

The ALICE O^2 reconstruction and analysis software

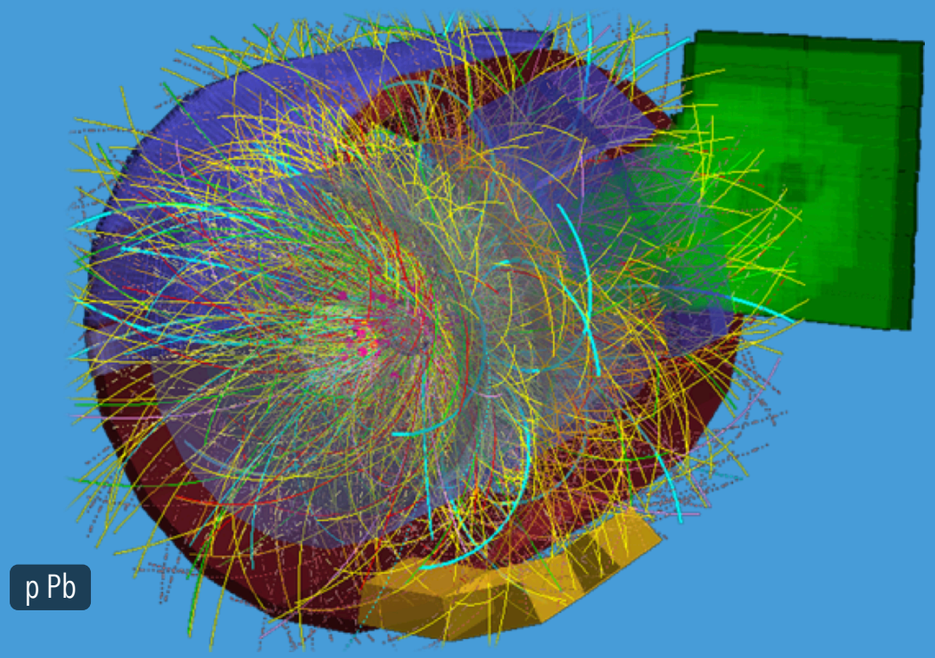
ALICE T1 / T2 Workshop, Seoul, Korea

Giulio Eulisse - CERN

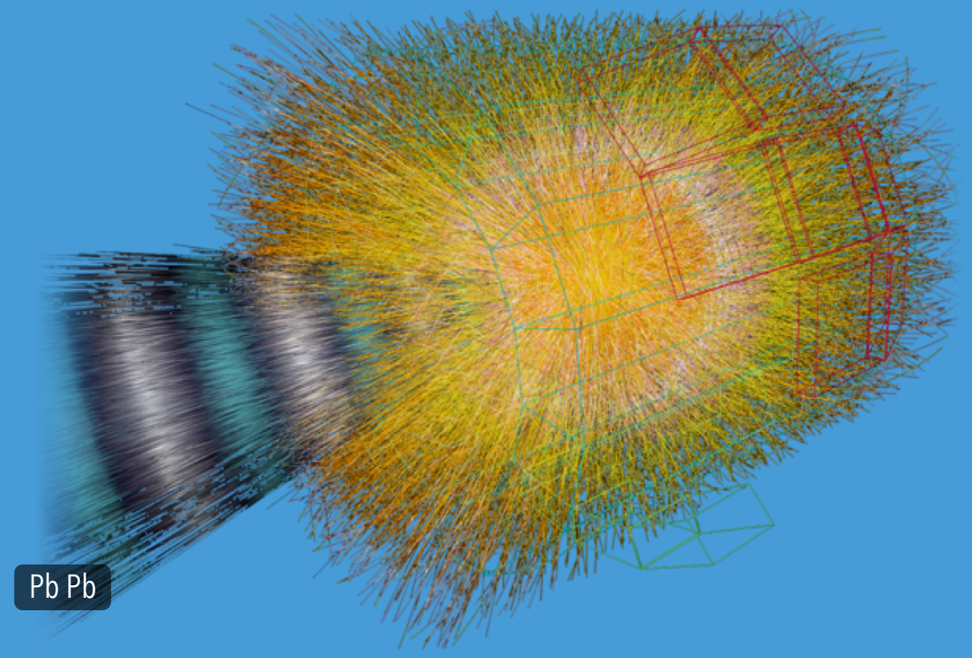
Run 2...



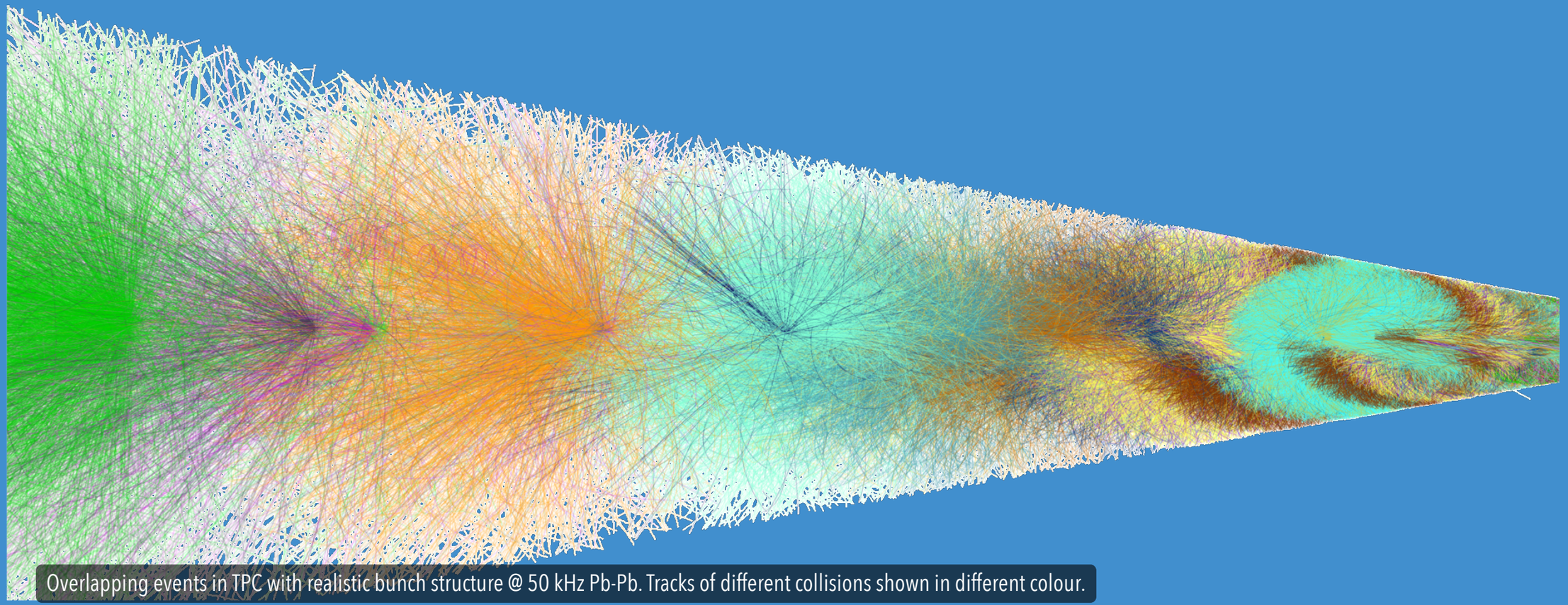
p b



p Pb



Pb Pb

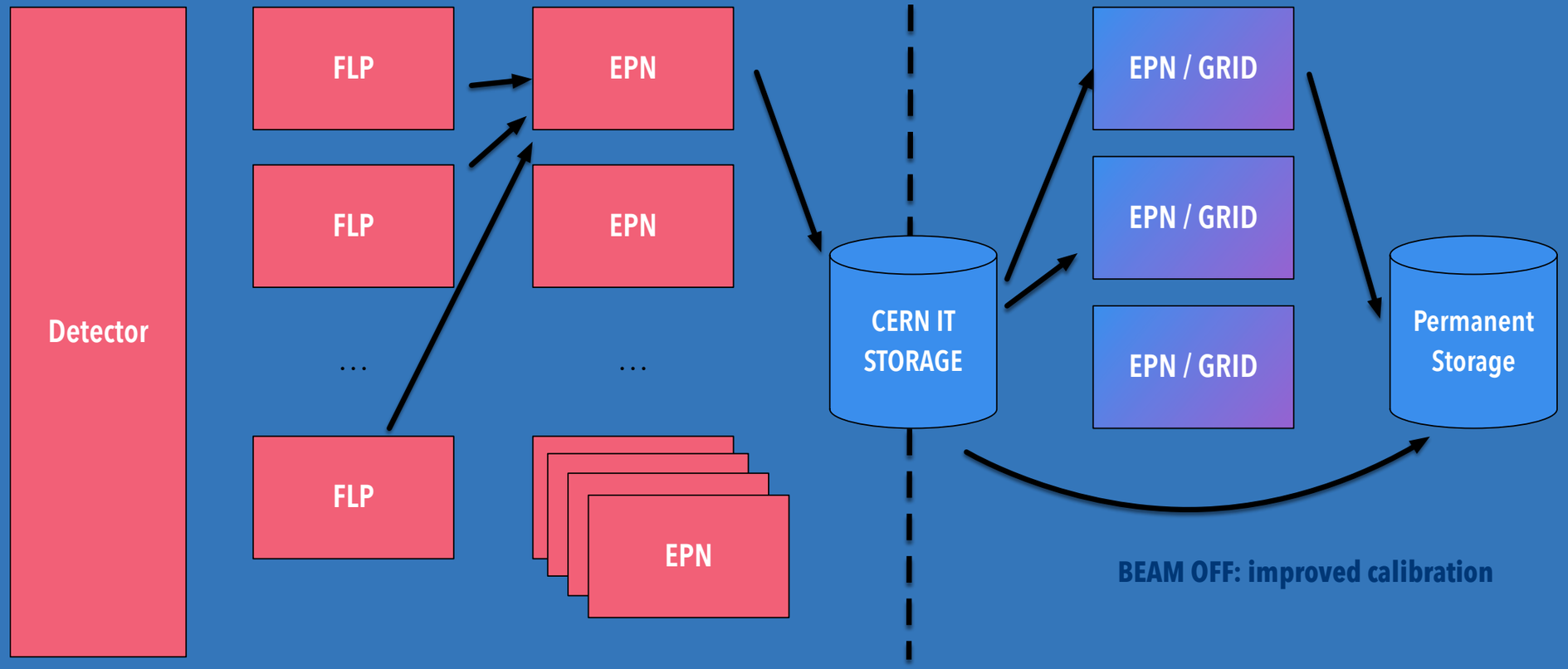


Overlapping events in TPC with realistic bunch structure @ 50 kHz Pb-Pb. Tracks of different collisions shown in different colour.

...to Run 3

O(100x) more events

Lossy online raw data compression GPUs!



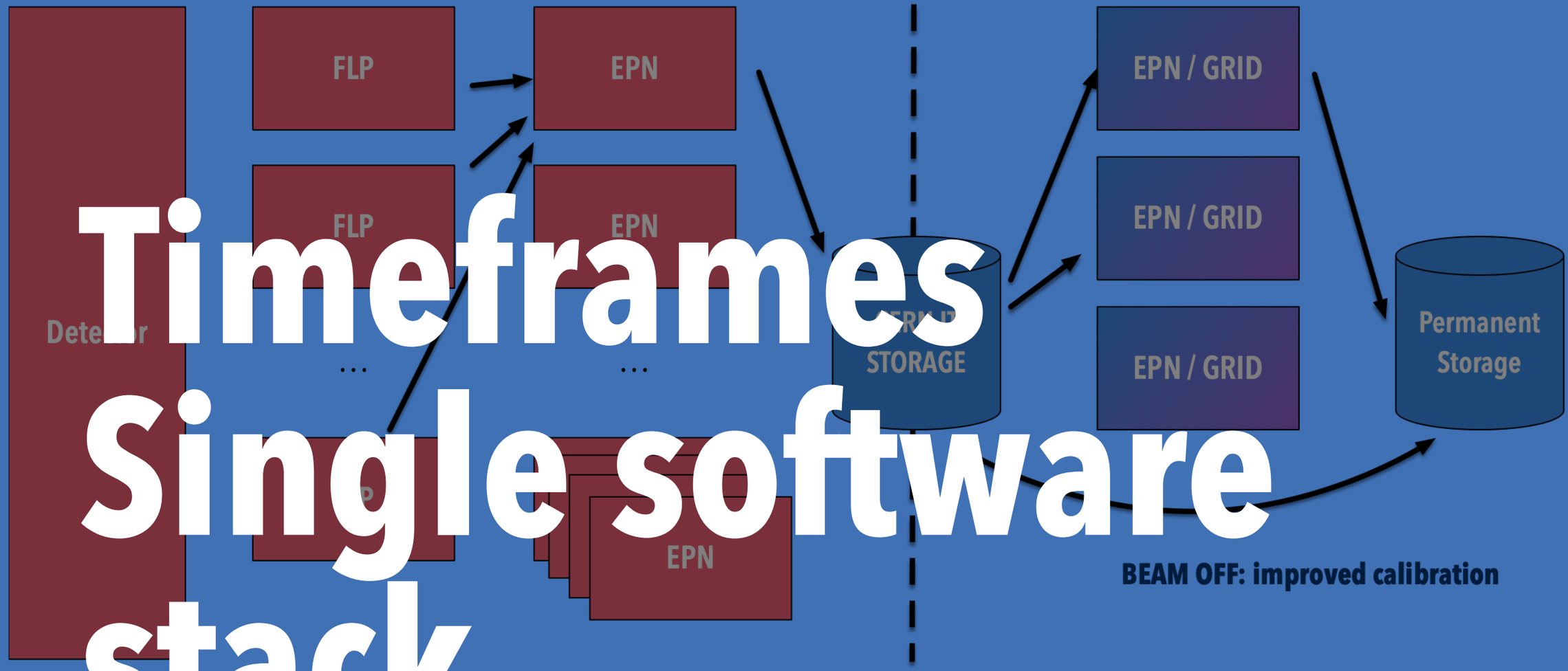
EPN Input data quantum is the
"timeframe": ~2.8 ms of continuous readout data. ~2.5 GB

BEAM ON: data reduction

Takeaway message:
one integrated system from data taking to final reconstruction (and beyond)

Timeframes

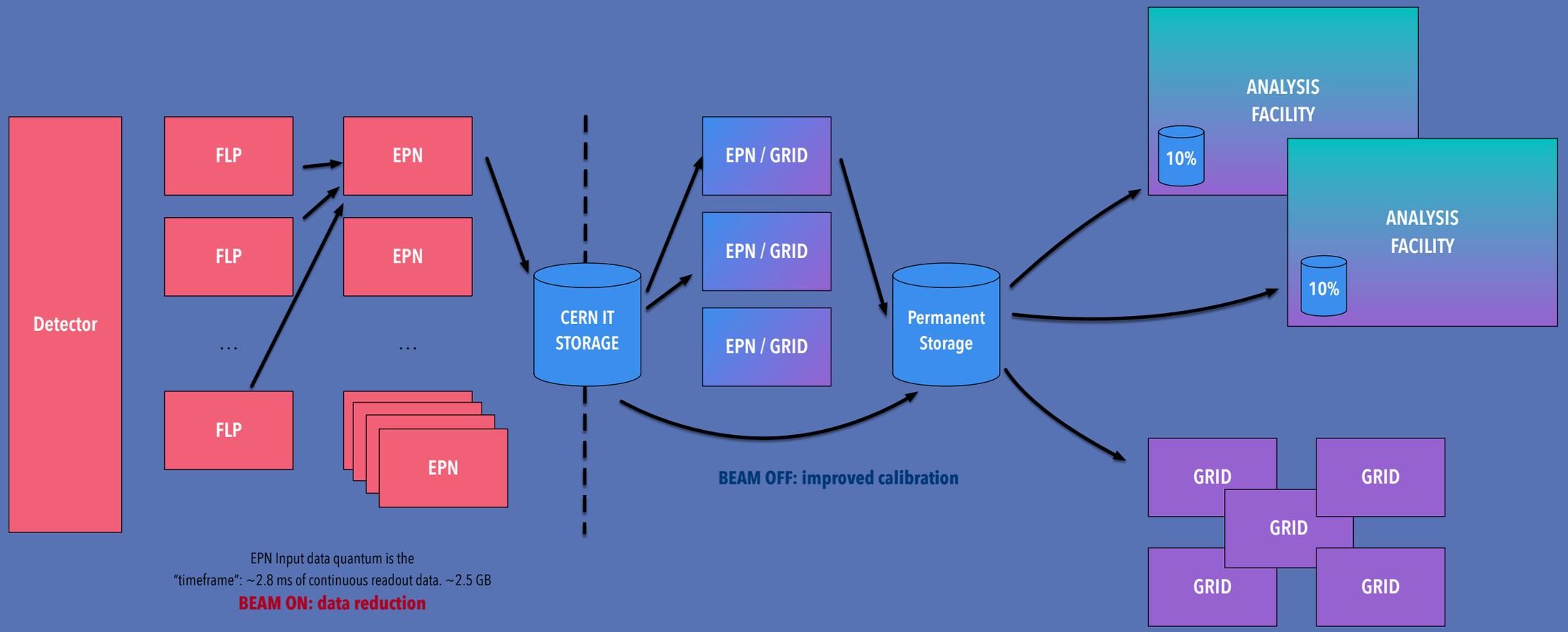
Single software stack

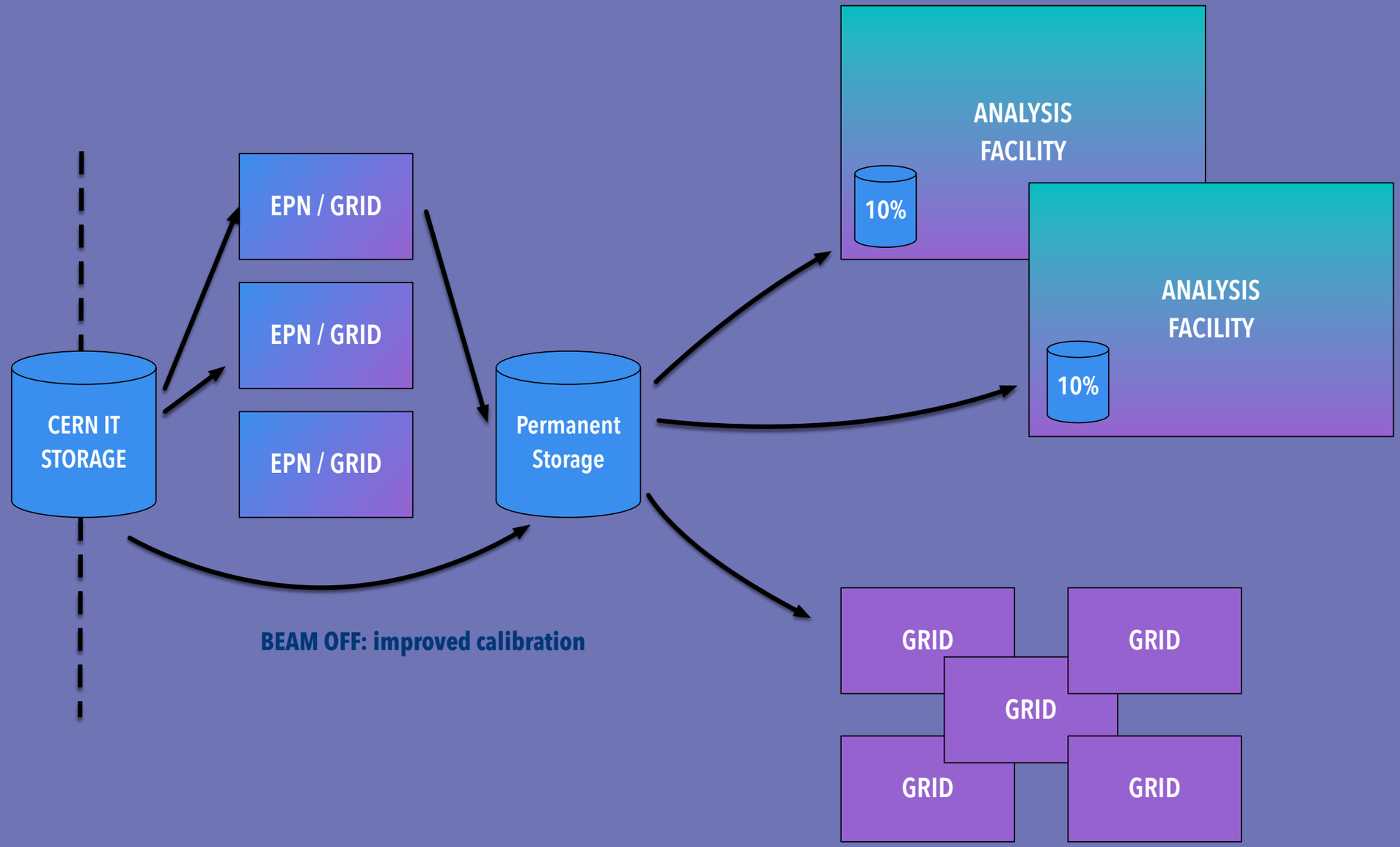


t data quantum is the
"timeframe": ~2.8 ms of continuous readout data. ~2.5 GB

BEAM ON: data reduction

Takeaway message:
one integrated system from data taking to final reconstruction (and beyond)


Detector
FLP
FLP
FLP
EPN
EPN
EPN
CERN IT STORAGE
EPN / GRID
EPN / GRID
EPN / GRID
Permanent Storage
ANALYSIS FACILITY
10%
ANALYSIS FACILITY
10%
GRID
GRID
GRID
GRID
GRID
EPN Input data quantum is the
"timeframe": ~2.8 ms of continuous readout data. ~2.5 GB
BEAM ON: data reduction
BEAM OFF: improved calibration


EPN / GRID
EPN / GRID
EPN / GRID
CERN IT STORAGE
Permanent Storage
ANALYSIS FACILITY
10%
ANALYSIS FACILITY
10%
BEAM OFF: improved calibration
GRID
GRID
GRID
GRID
GRID

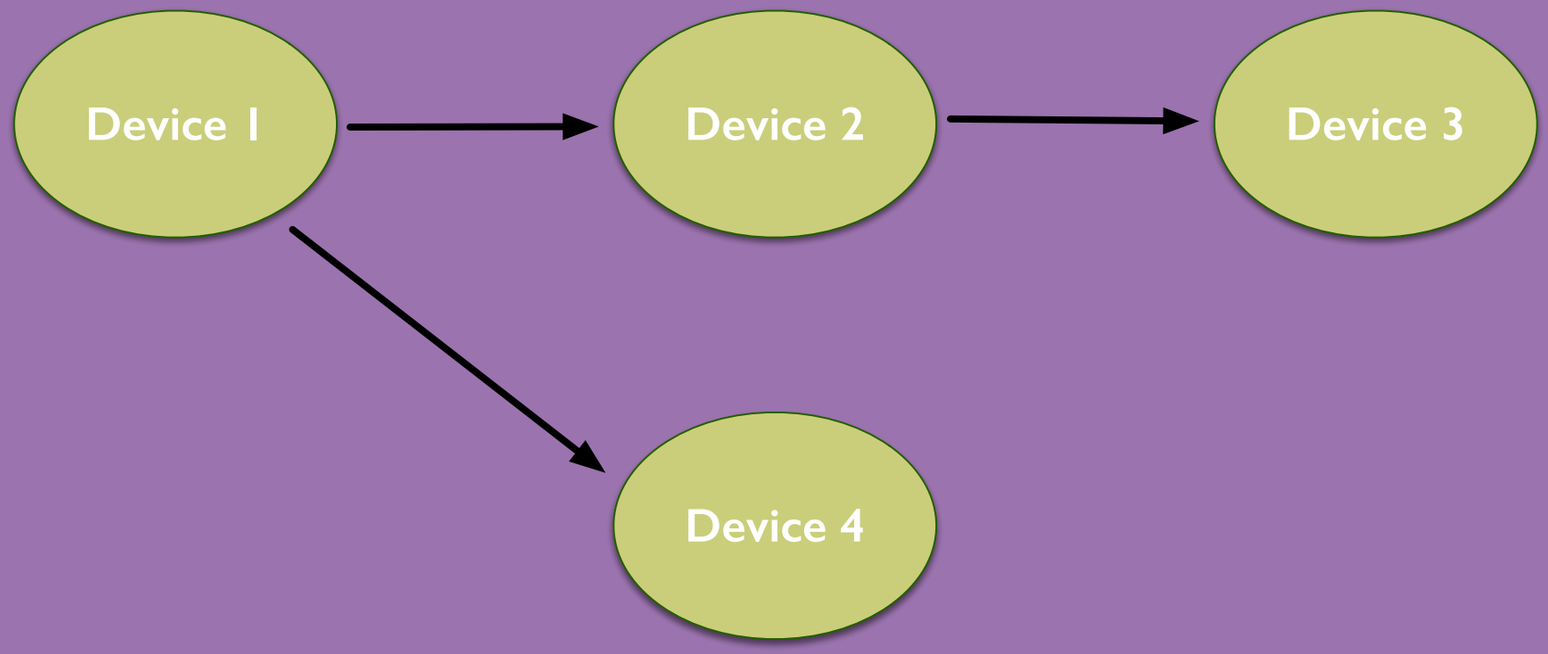
현위치

CERN ❤️ GSI

Actor Model

Shared memory optimised

Transport Layer: ALFA / FairMQ[1]



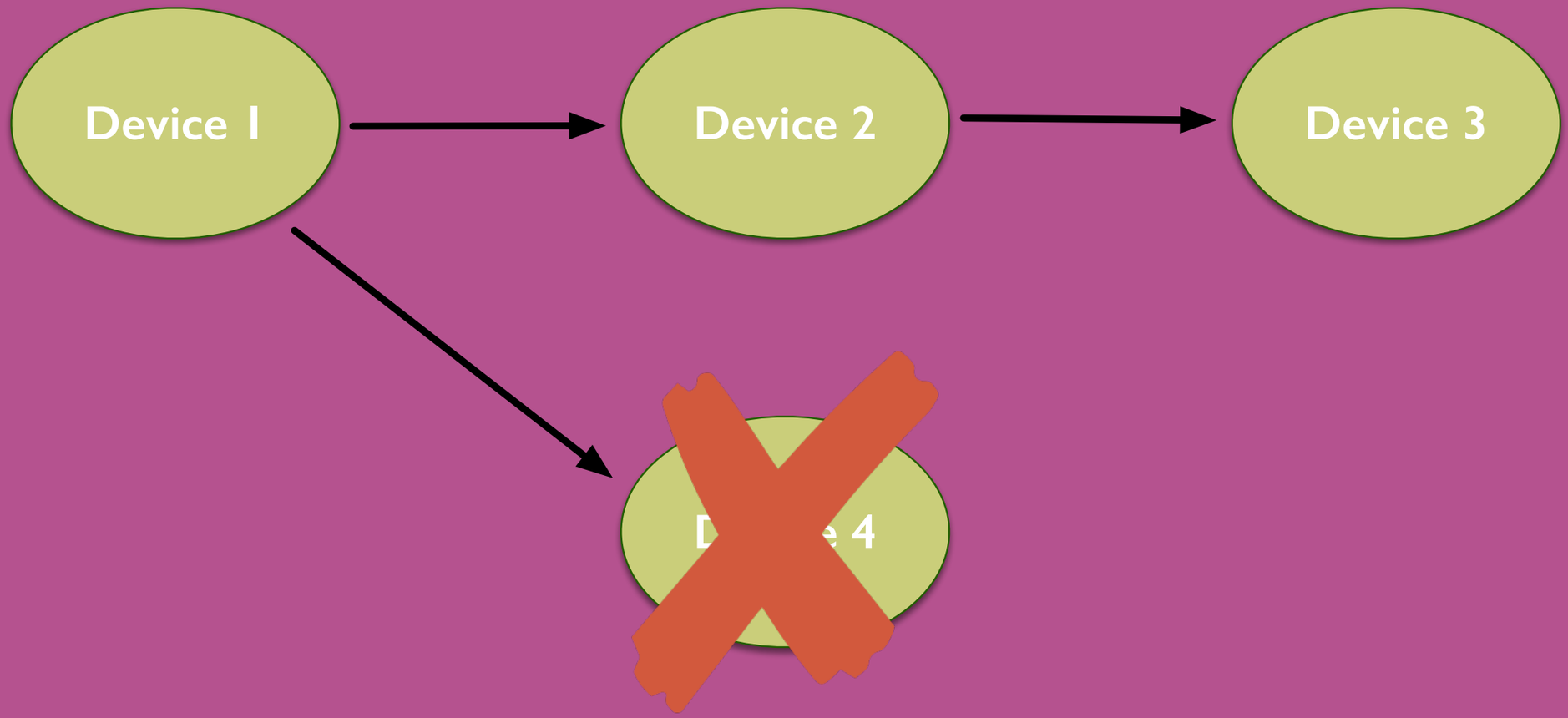
Devices, topologies & channels

one executable to many devices

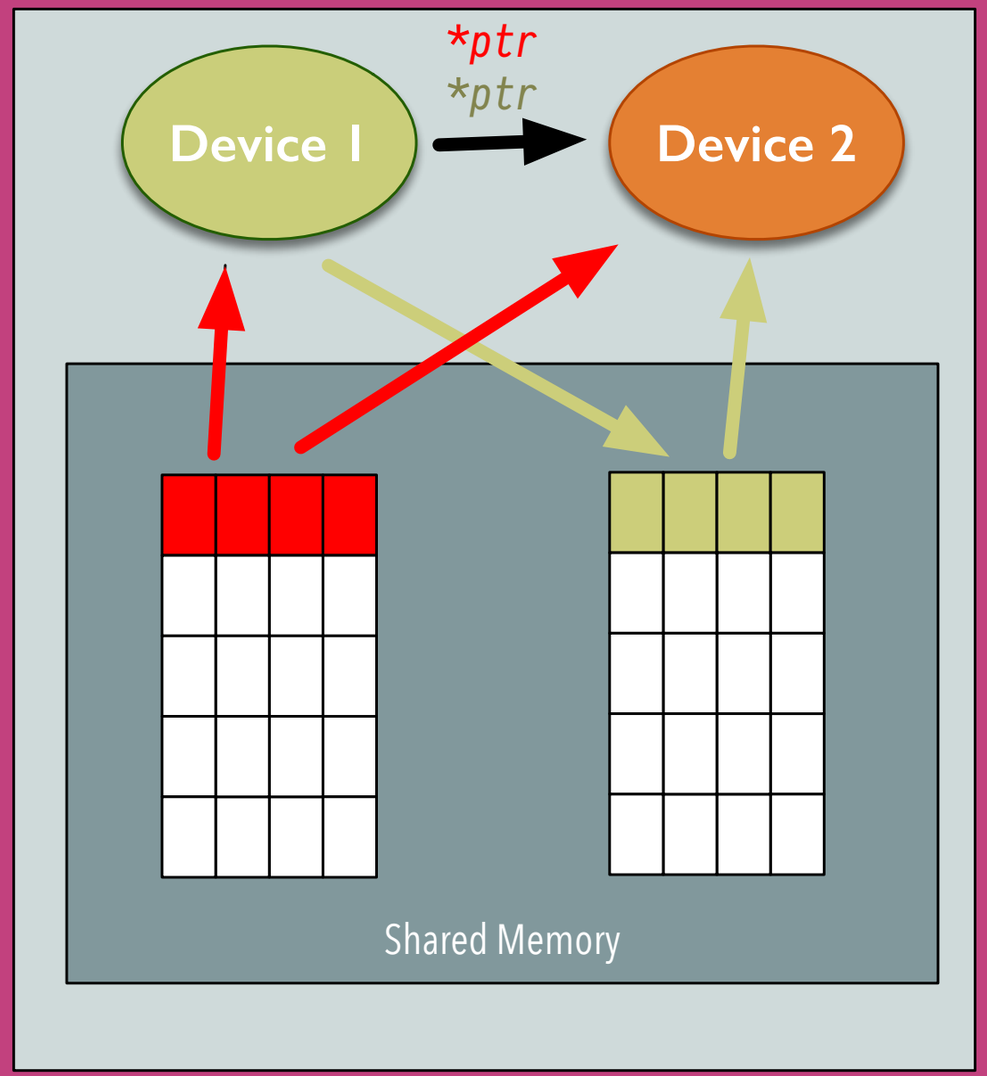
```
> ps aux | grep -e "^o2-"
```

--id & _tX suffix

```
o2-my-analysis-workflow --id <something> ...
```



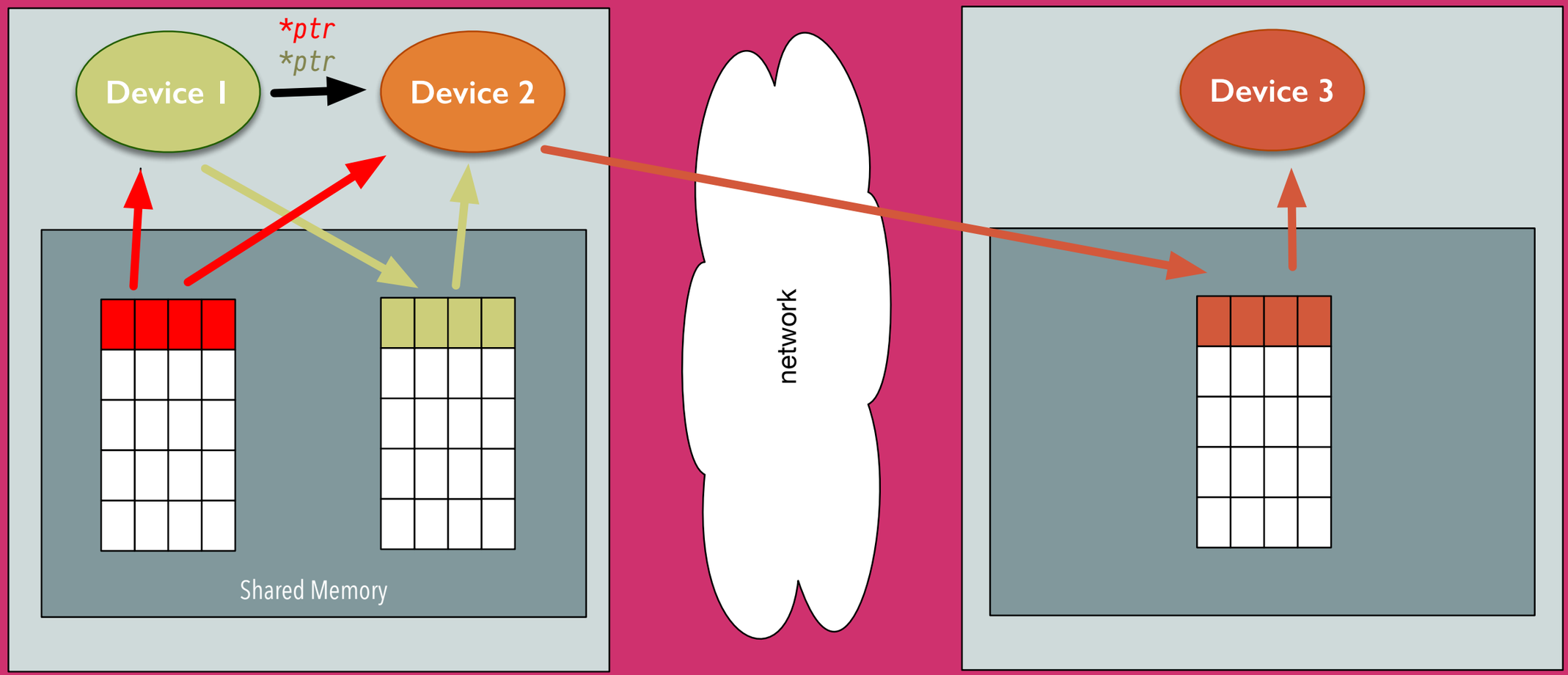
"expendable"

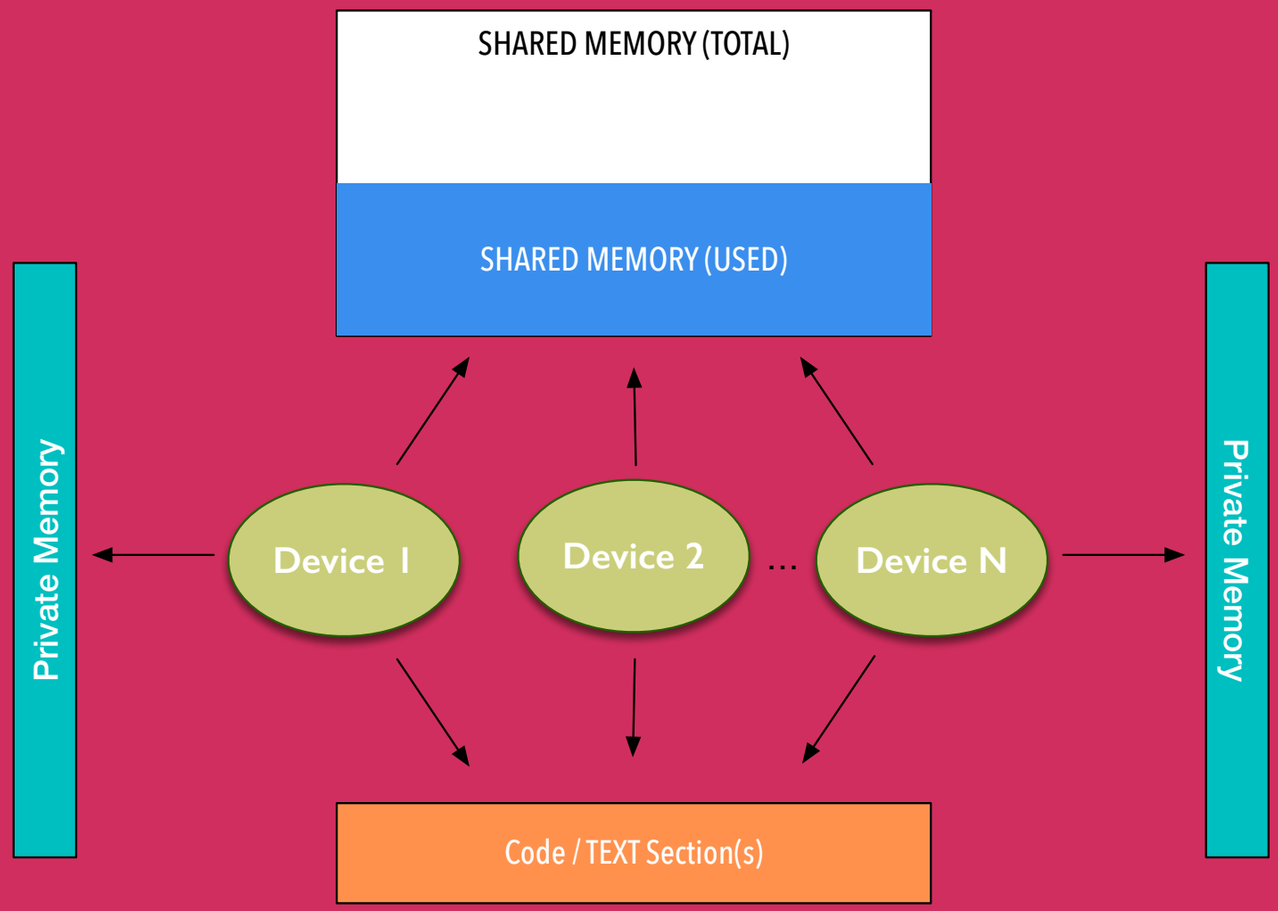


boost::interprocess + zeromq

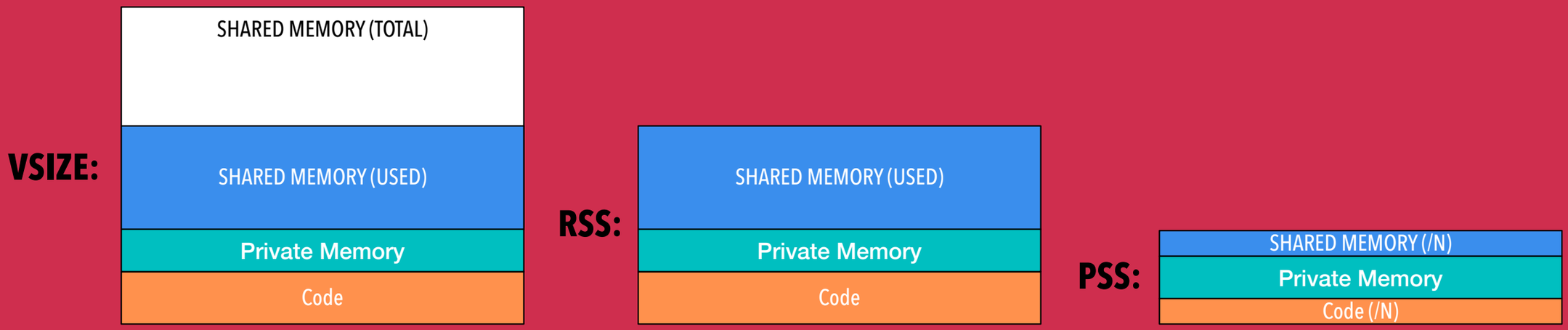
/dev/shm & fairmq-shmmonitor

```
--session <id>
```


*ptr
*ptr
Device 1
Device 2
Shared Memory
network
Device 3



About memory accounting...



USE PSS as a memory metric!

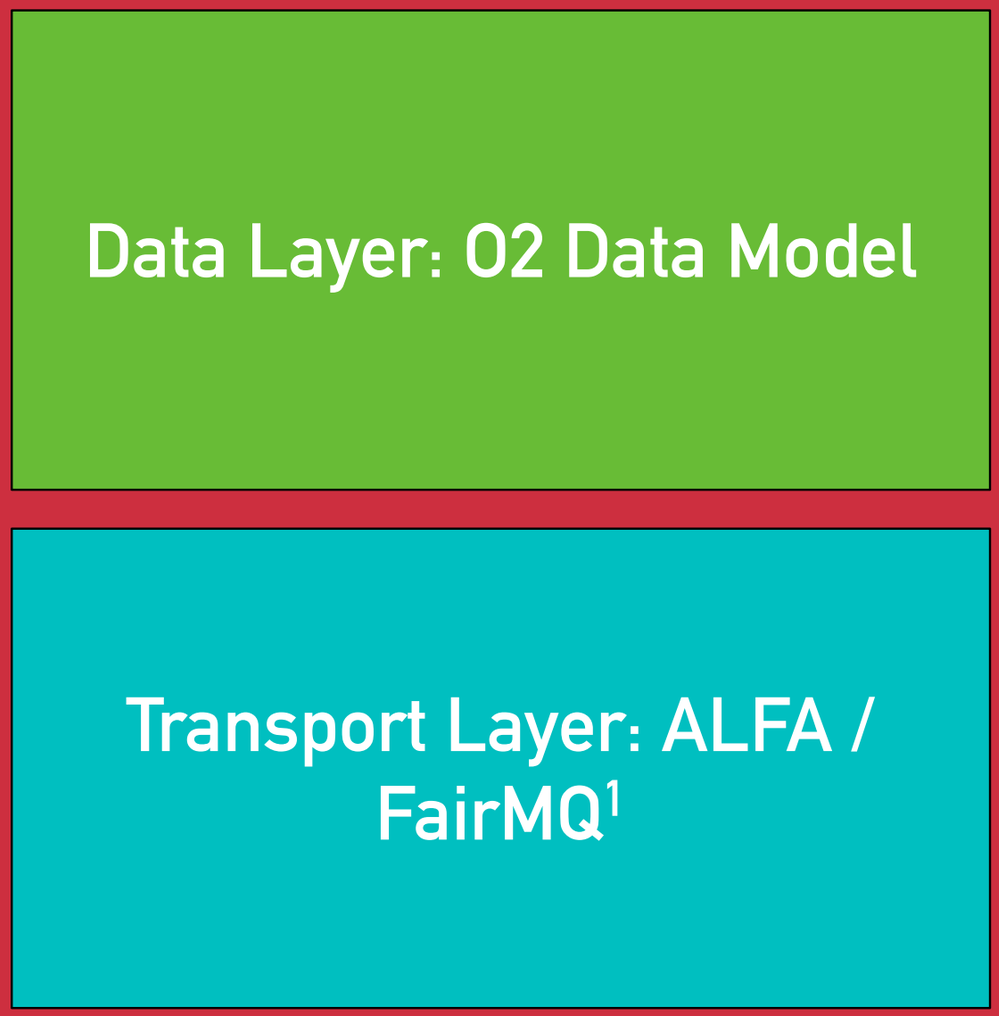
Shared memory at exit

MADV_PAGEOUT (linux 5.4)

Data layer

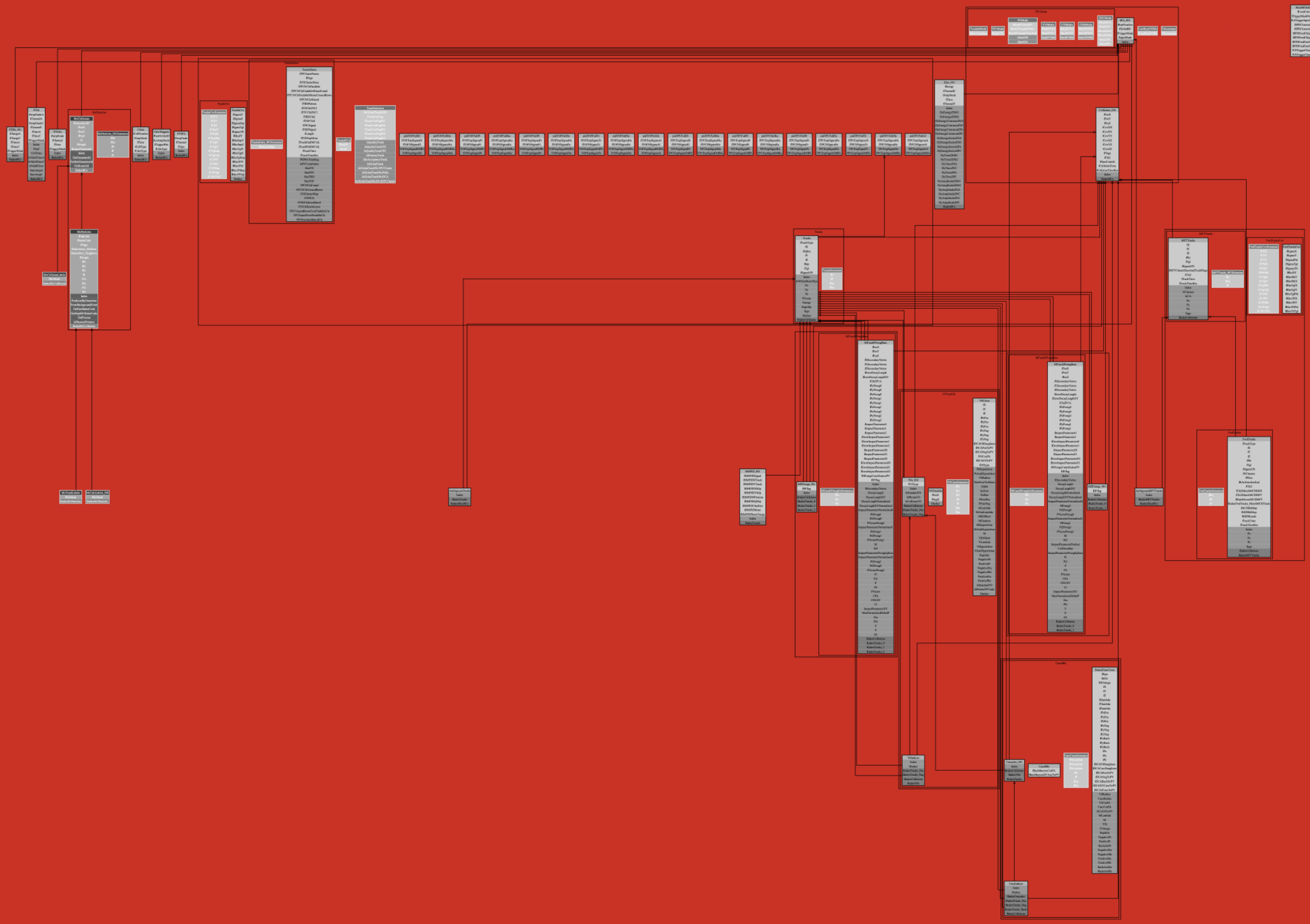
Uniform & predictable

Multibackend



Think about an in-memory per-(few)-timeframes database.

O^2 Analysis Data Model



Relational DB like

Bulk operations

```
track.pt()
```

Apache Arrow

X	Y	Z	1/Pt	E

Only two hard problems in distributed systems

2. Exactly-once delivery

1. Guaranteed order of messages

2. Exactly-once delivery

Framework &
Data Processing Layer (DPL)

Data Layer: O2 Data Model

Transport Layer: ALFA / FairMQ[1]

Simplify

Data Flow system

Implicit topology definition

Common services

Access to conditions

Access to files

Plugin manager

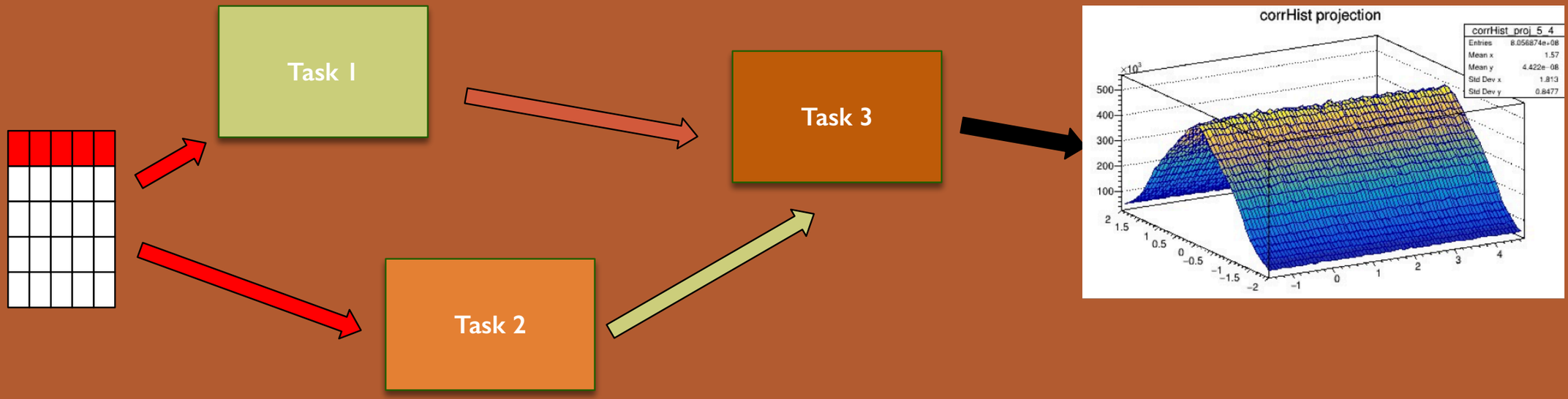
Debug tools

Integration

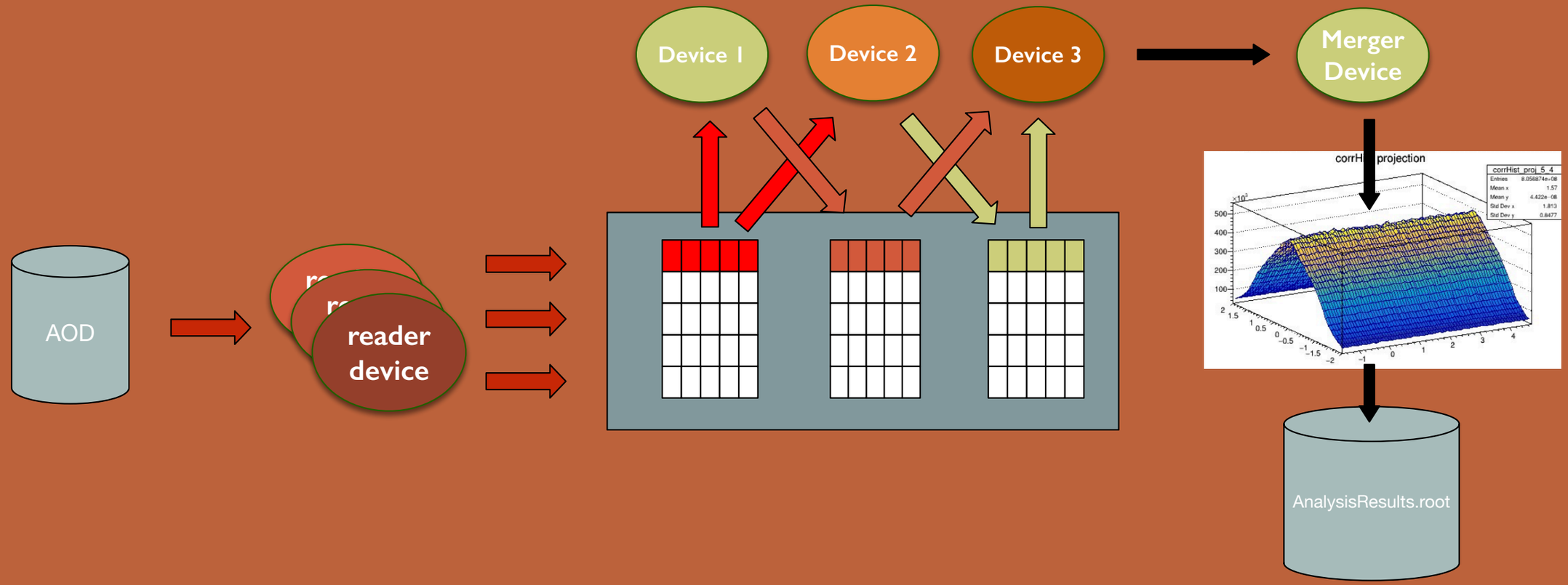
Monitoring

Logging

Control



Users provides tasks...



...DPL builds a topology for them

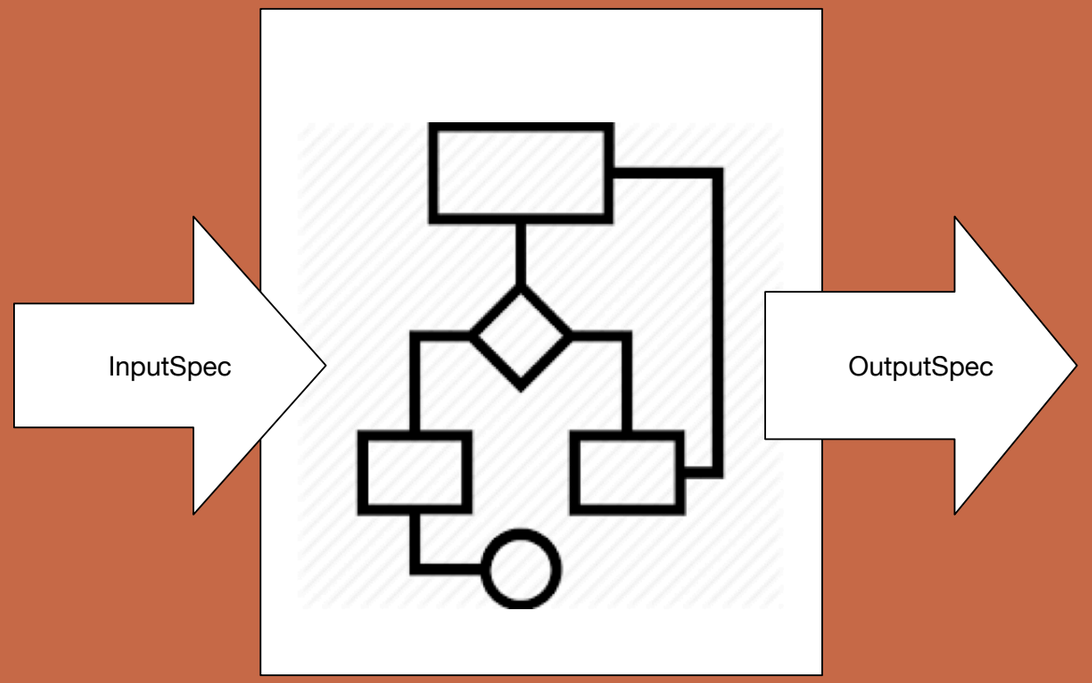
Data Processing Layer: Building block

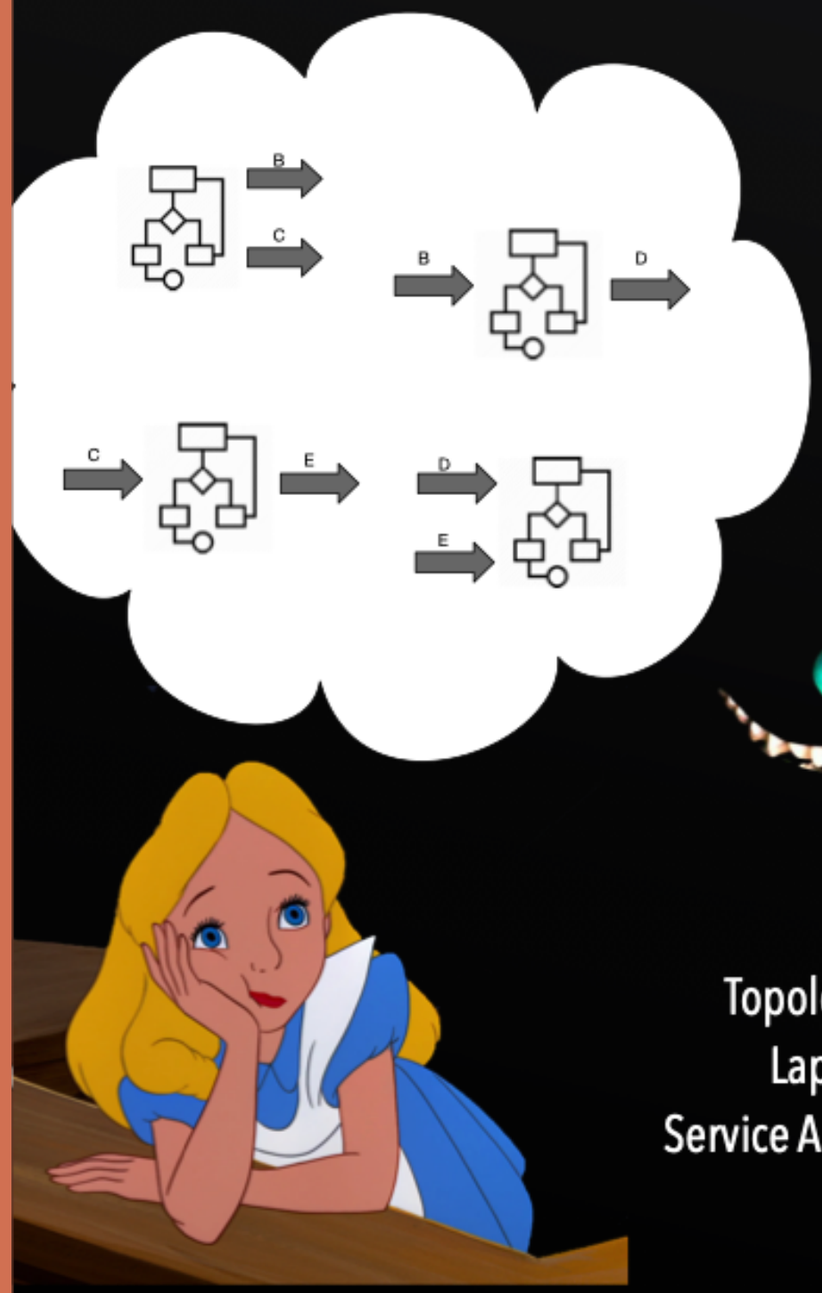
A DataProcessorSpec defines a pipeline stage as a building block.

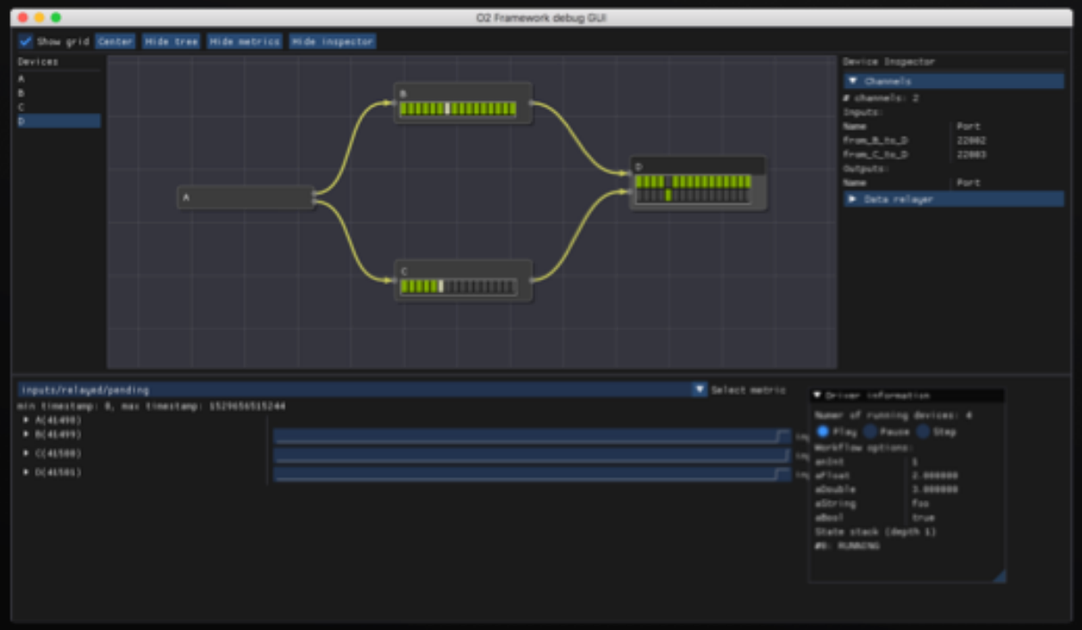
Specifies inputs and outputs in terms of the O2 Data Model descriptors.

Provide an implementation of how to act on the inputs to produce the output.

Advanced user can express possible data or time parallelism opportunities.





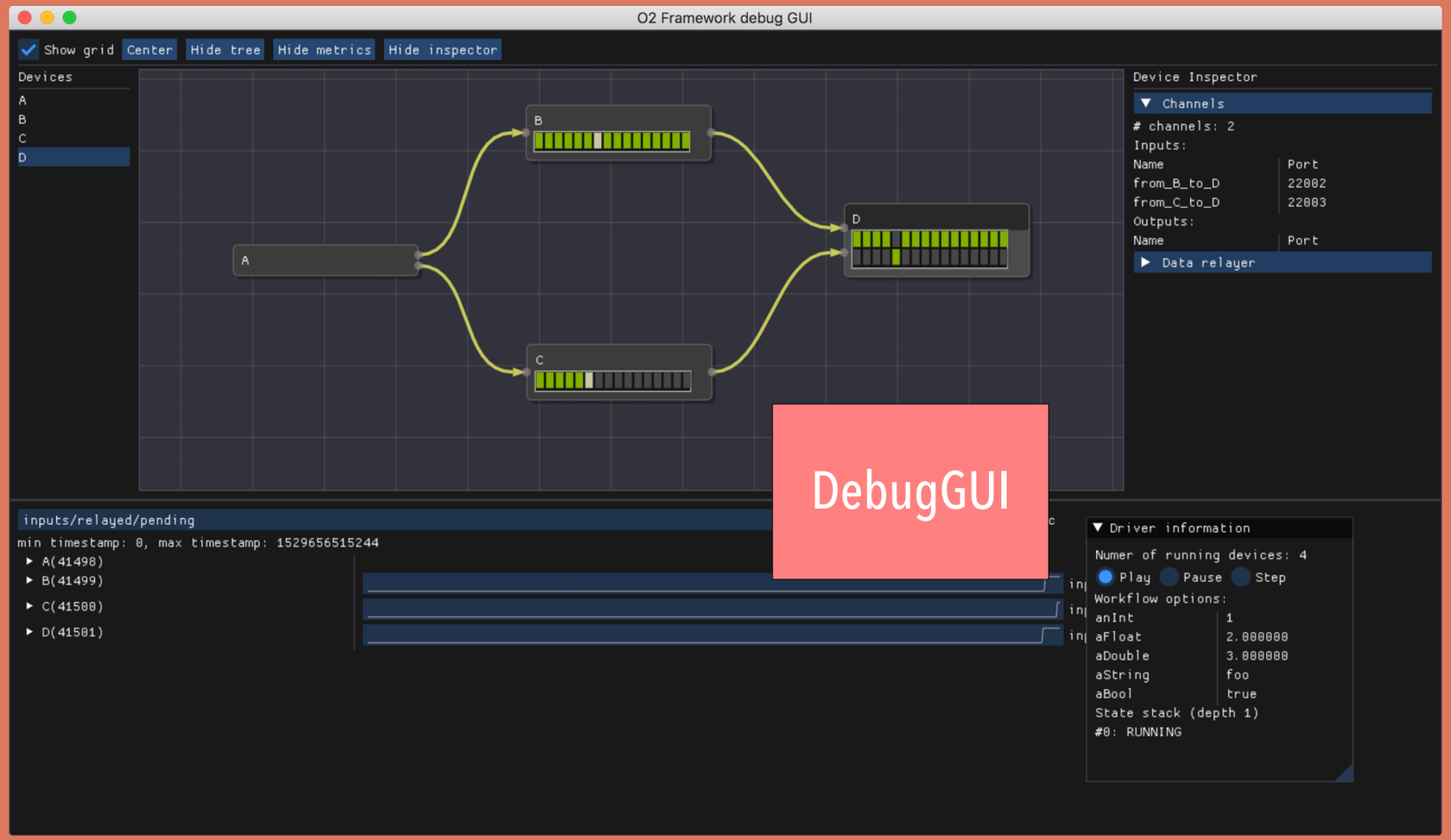


Topology is defined implicitly.

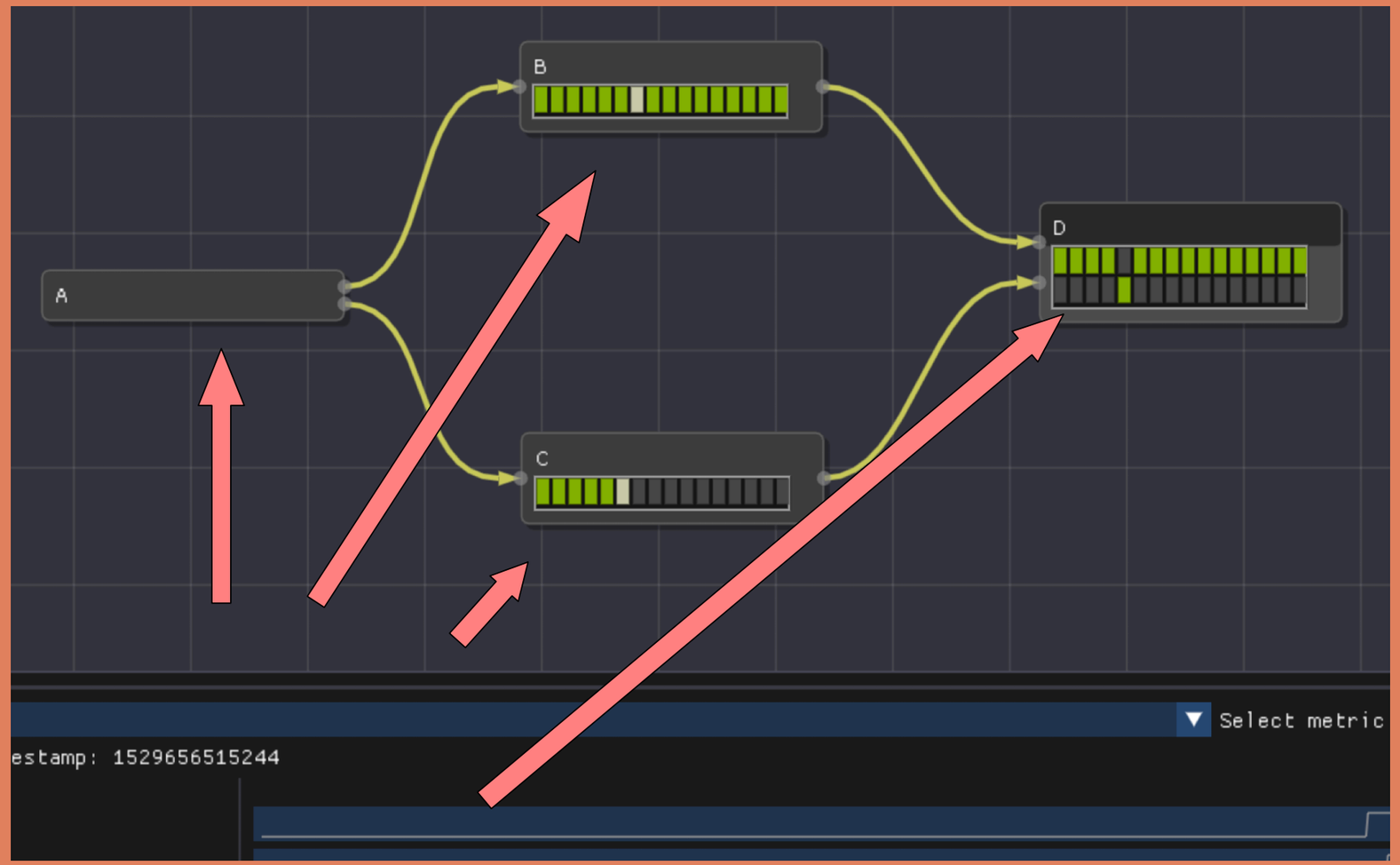
Topological sort ensures a viable dataflow is constructed (no cycles!).

Laptop users gets immediate feedback through the debug GUI.

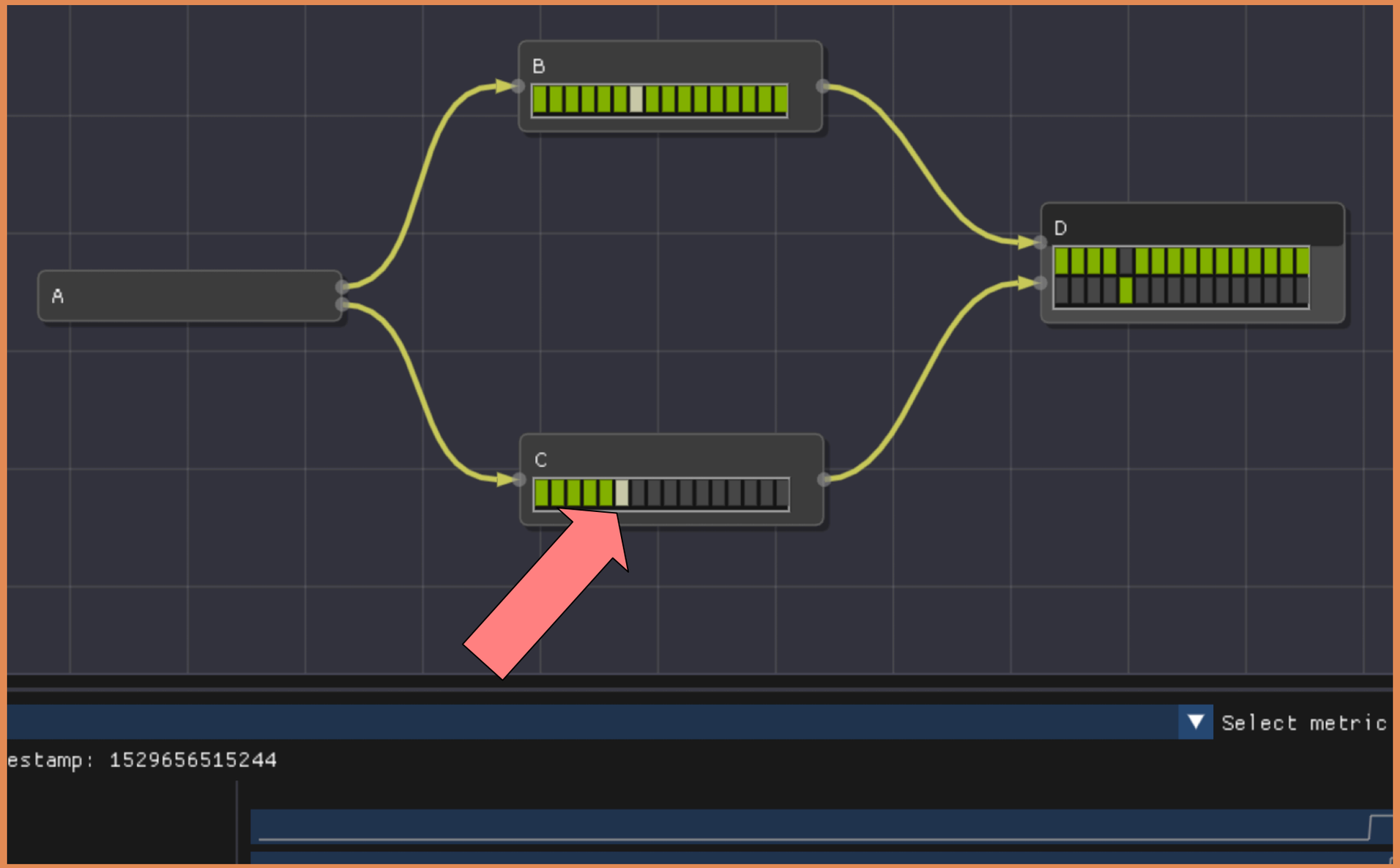
Service API allows integration with non data flow components (e.g. Control)


O2 Framework debug GUI
Show grid
Center
Hide tree
Hide metrics
Hide inspector
Devices
A
B
C
D
A
B
C
D
Device Inspector
Channels
channels: 2
Inputs:
Name
Port
from_B_to_D
22002
from_C_to_D
22003
Outputs:
Name
Port
Data relayer
DebugGUI
inputs/relayed/pending
min timestamp: 0, max timestamp: 1529656515244
A(41498)
B(41499)
C(41500)
D(41501)
Driver information
Numer of running devices: 4
Play
Pause
Step
Workflow options:
anInt
1
aFloat
2.000000
aDouble
3.000000
aString
foo
aBool
true
State stack (depth 1)
#0: RUNNING

DebugGUI

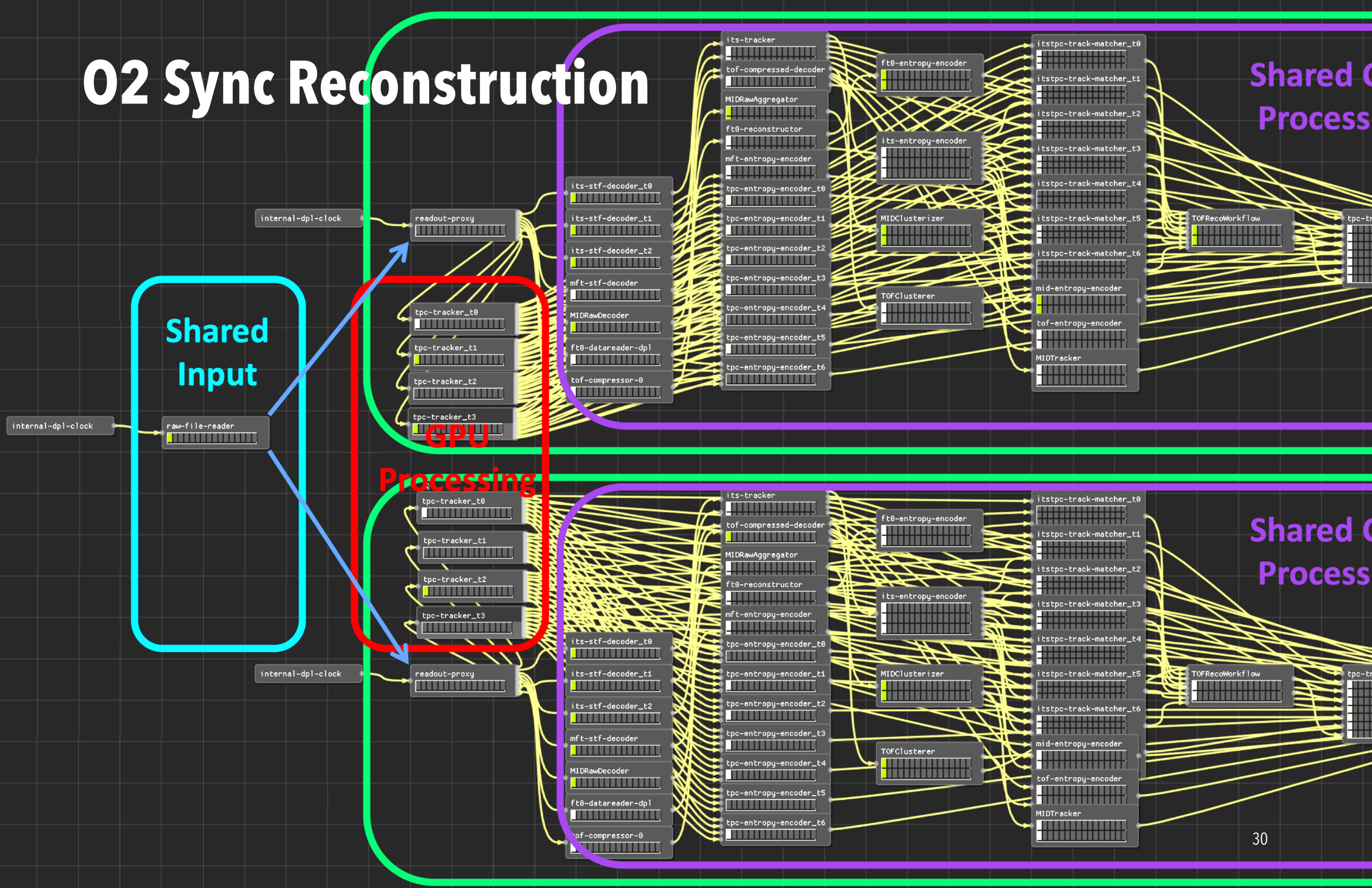


Each box represents a device, arrows

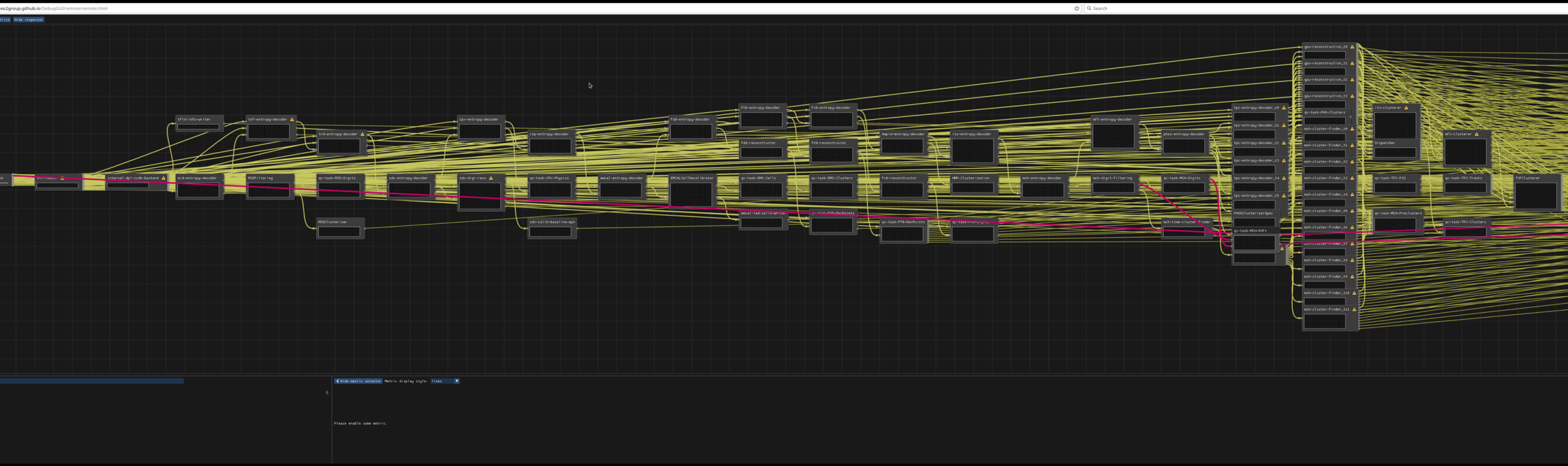


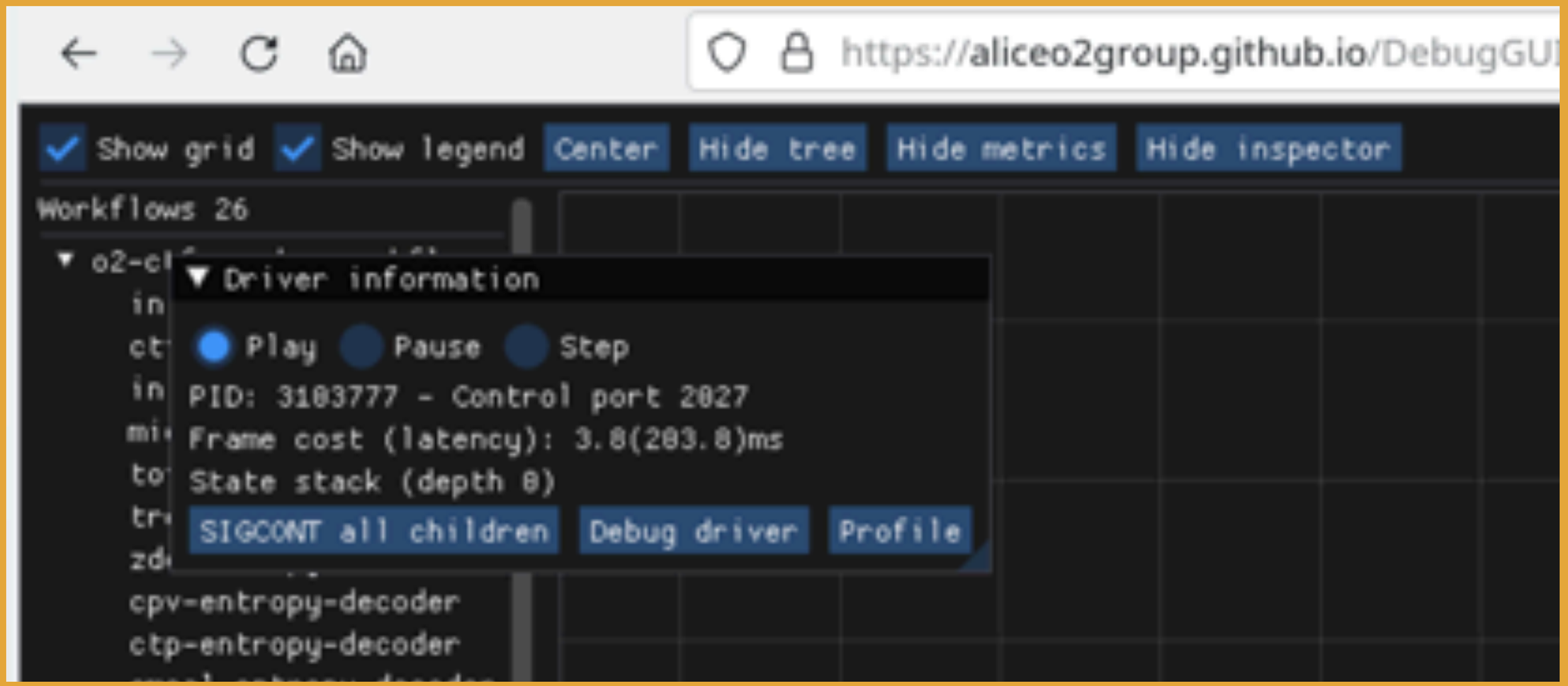
Each green lighted square is a message received

O2 Sync Reconstruction

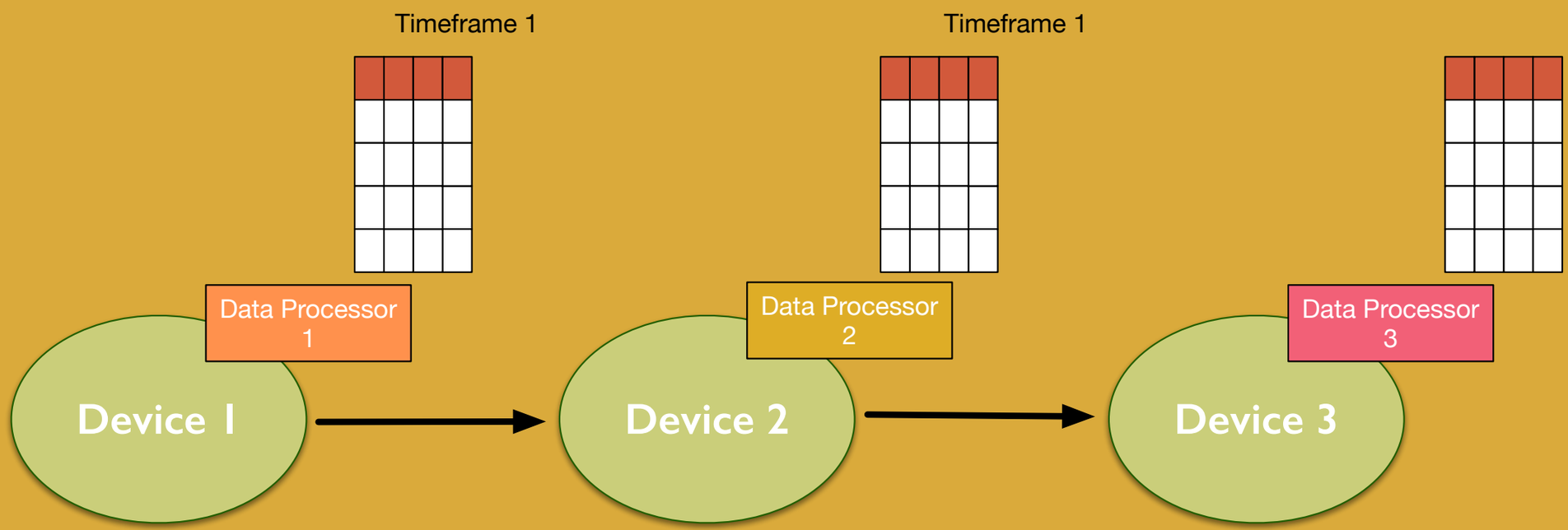


Async Reconstruction

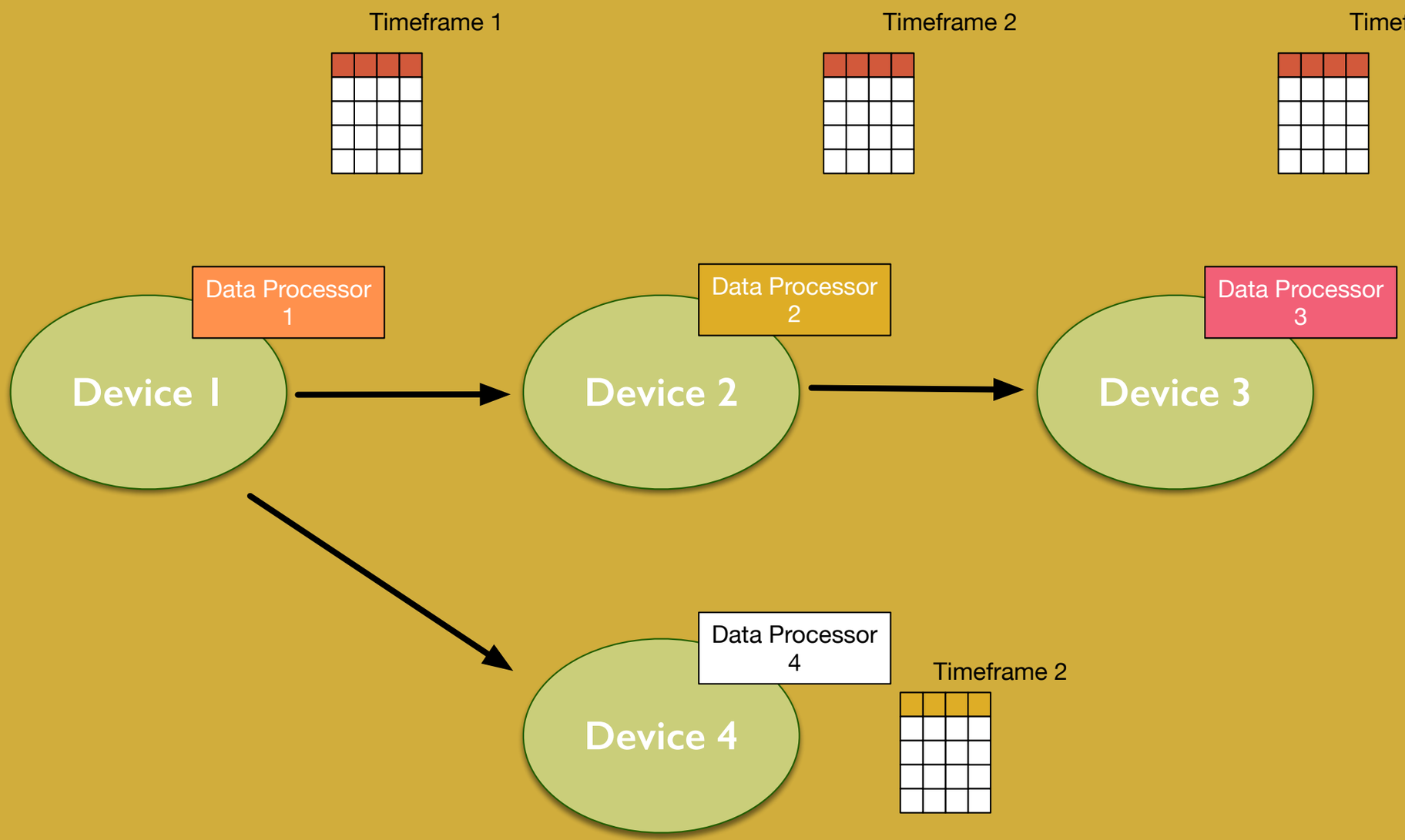



https://aliceo2group.github.io/DebugGUI
Show grid
Show legend
Center
Hide tree
Hide metrics
Hide inspector
Workflows 26
Driver information
Play
Pause
Step
PID: 3103777 - Control port 2027
Frame cost (latency): 3.8(203.8)ms
State stack (depth 0)
SIGCONT all children
Debug driver
Profile
cpv-entropy-decoder
ctp-entropy-decoder

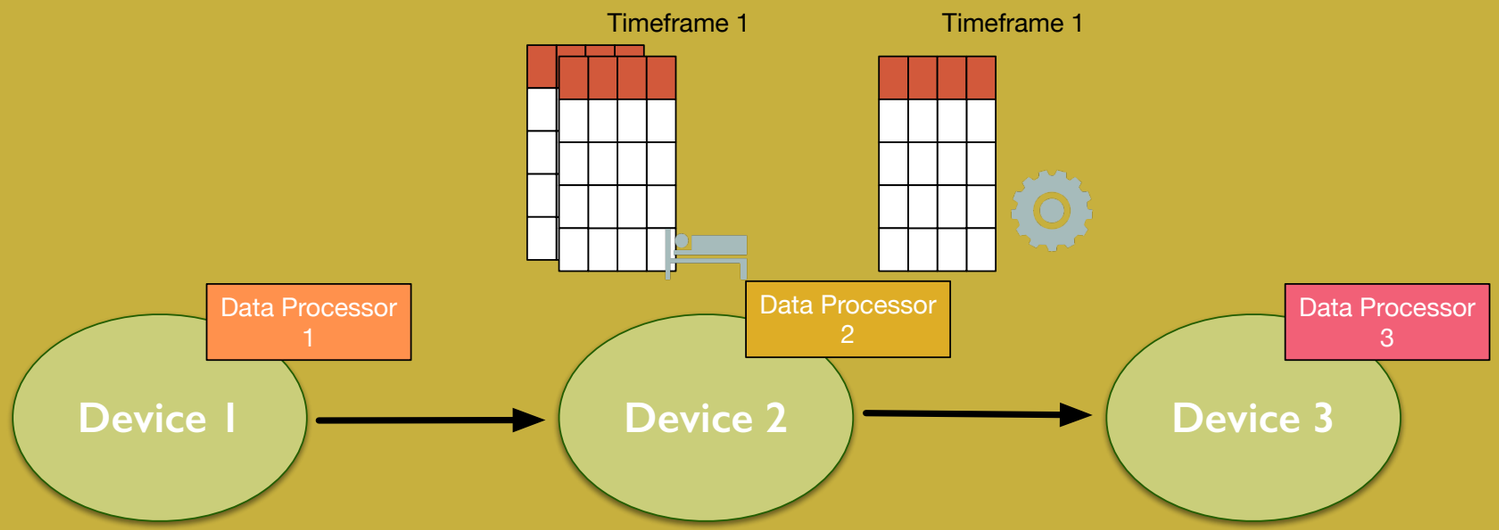
Remote!



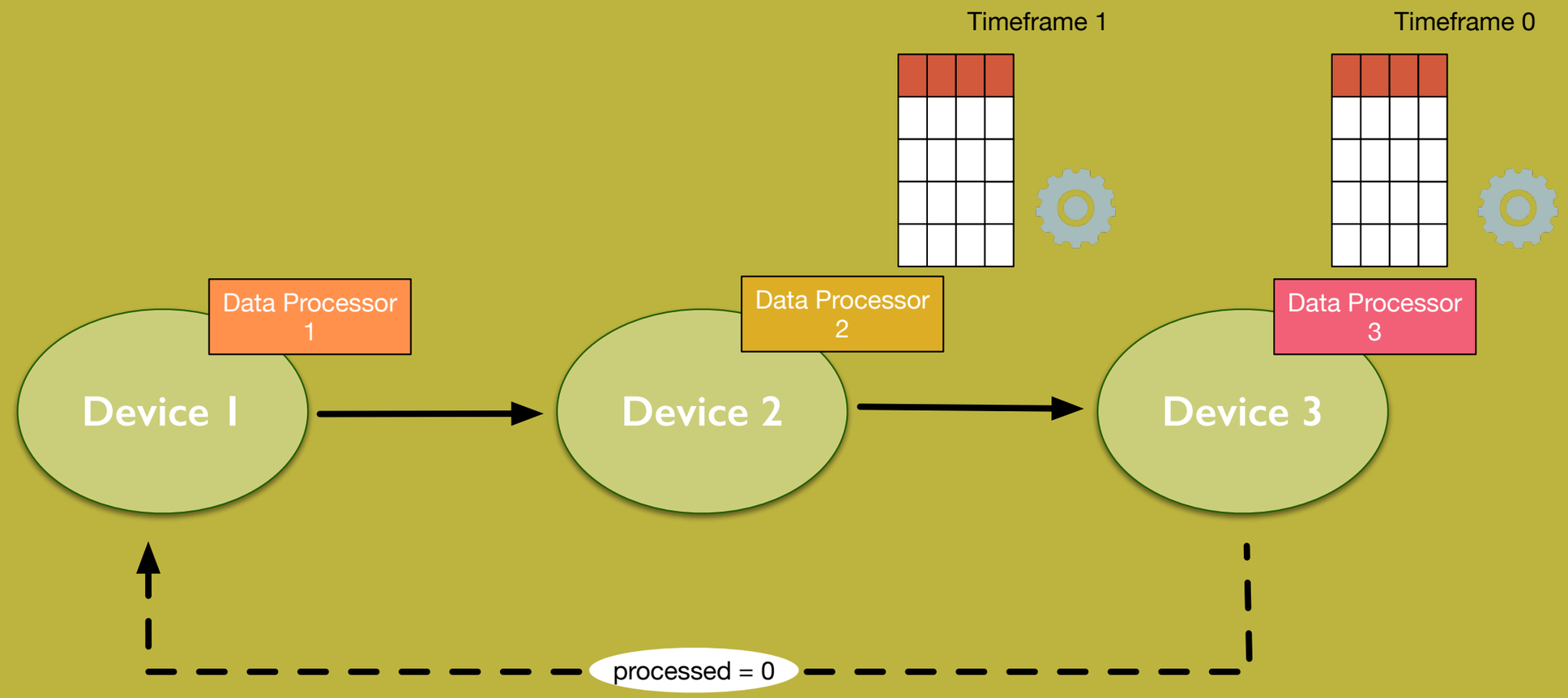
natural parallelism



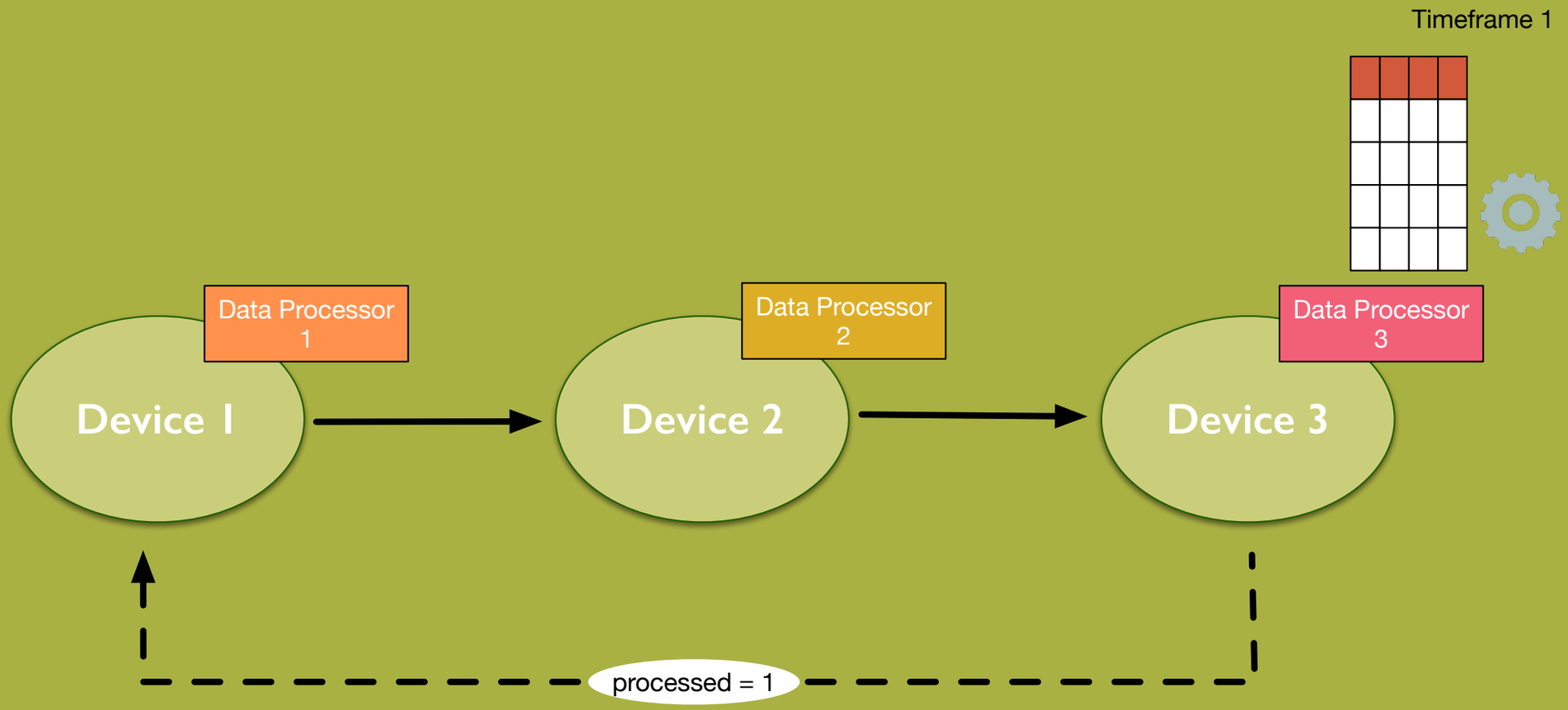
vertical parallelism



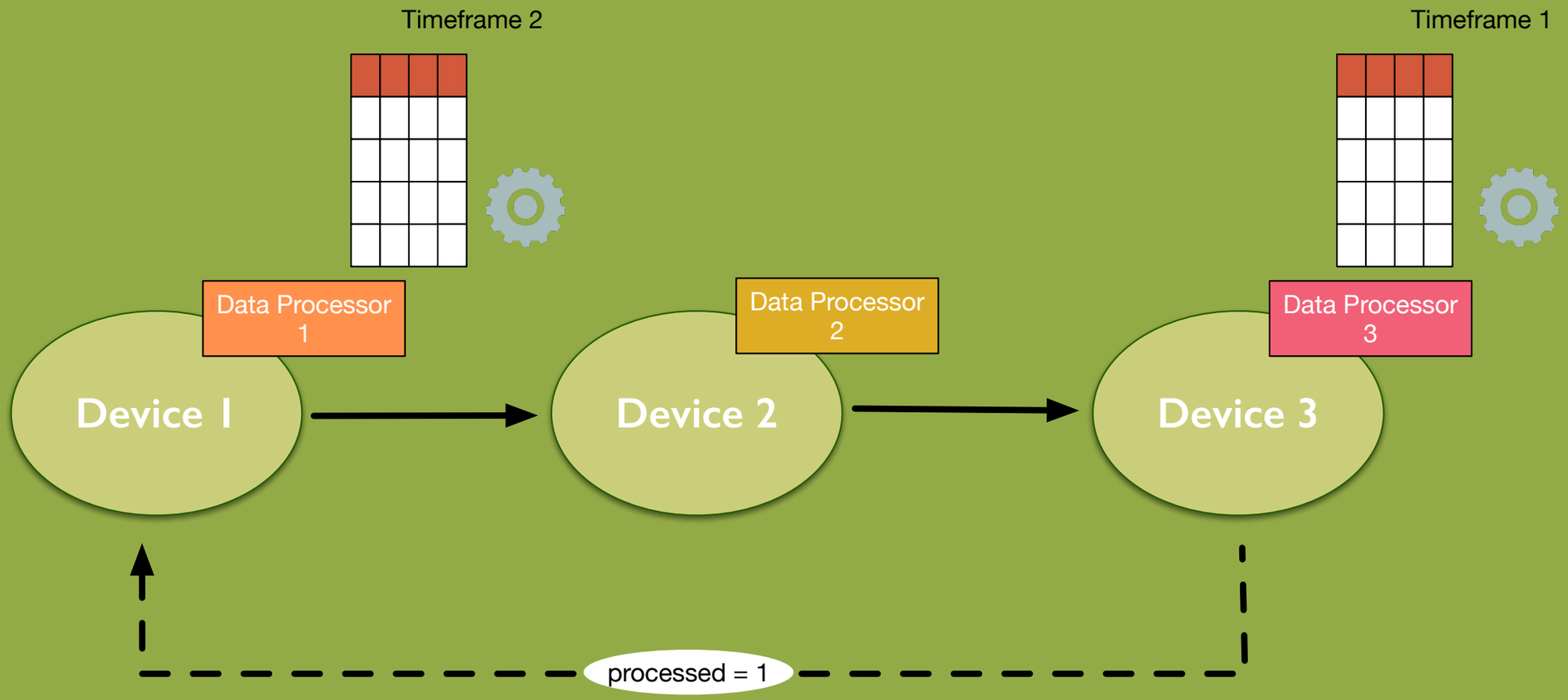
Rate Limiting to the rescue!



back channel to report progress



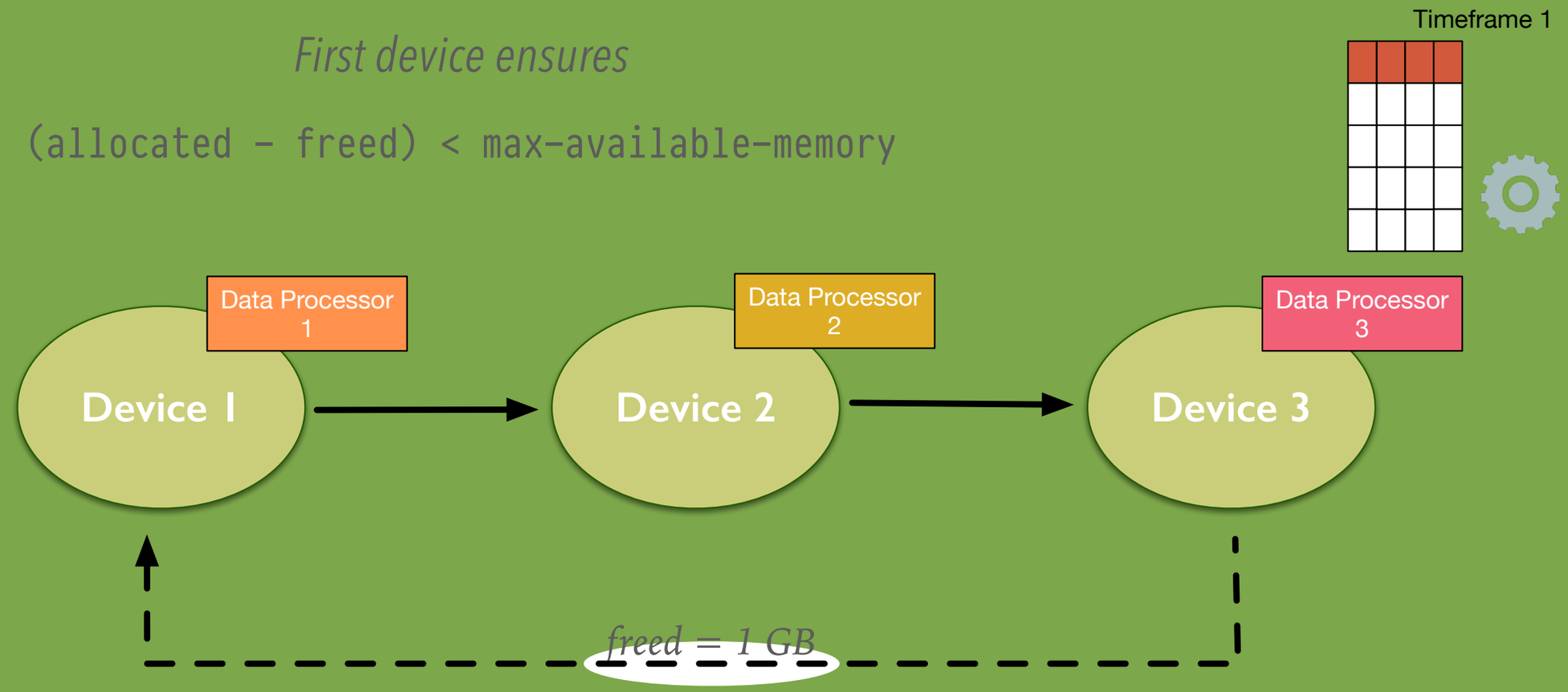
sink reports processed timeframes



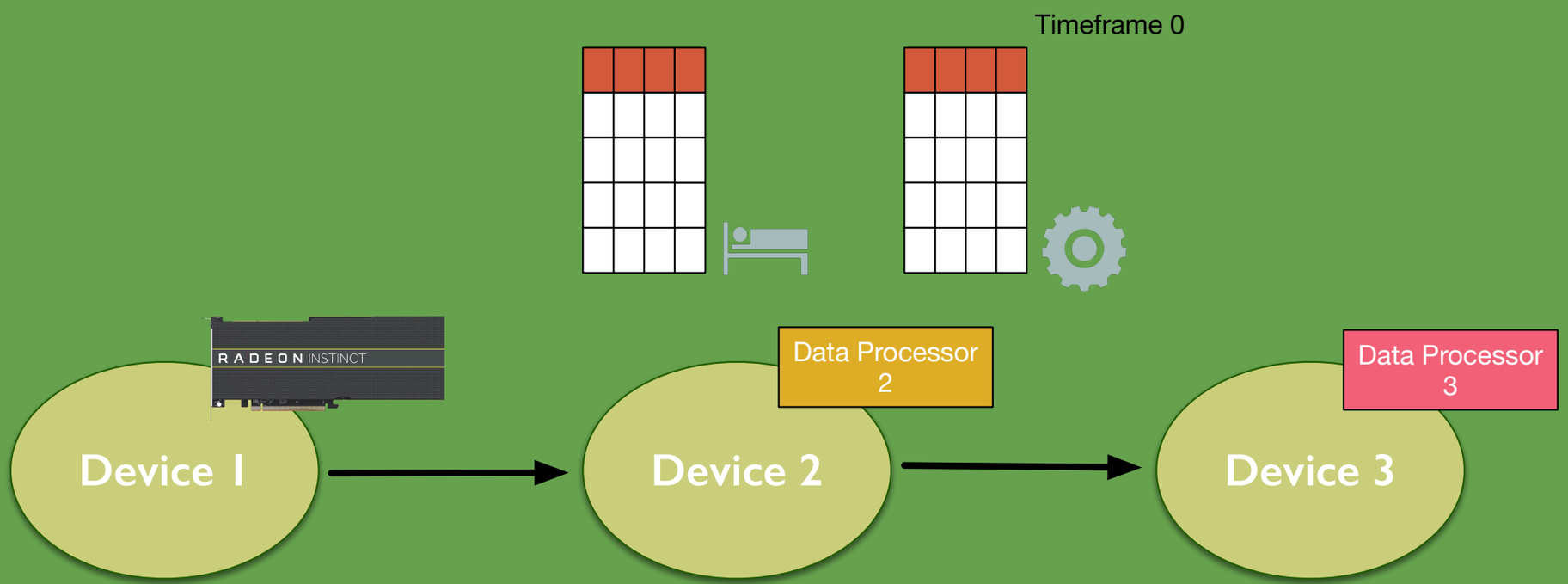
...first device injects a new timeframe.

First device ensures

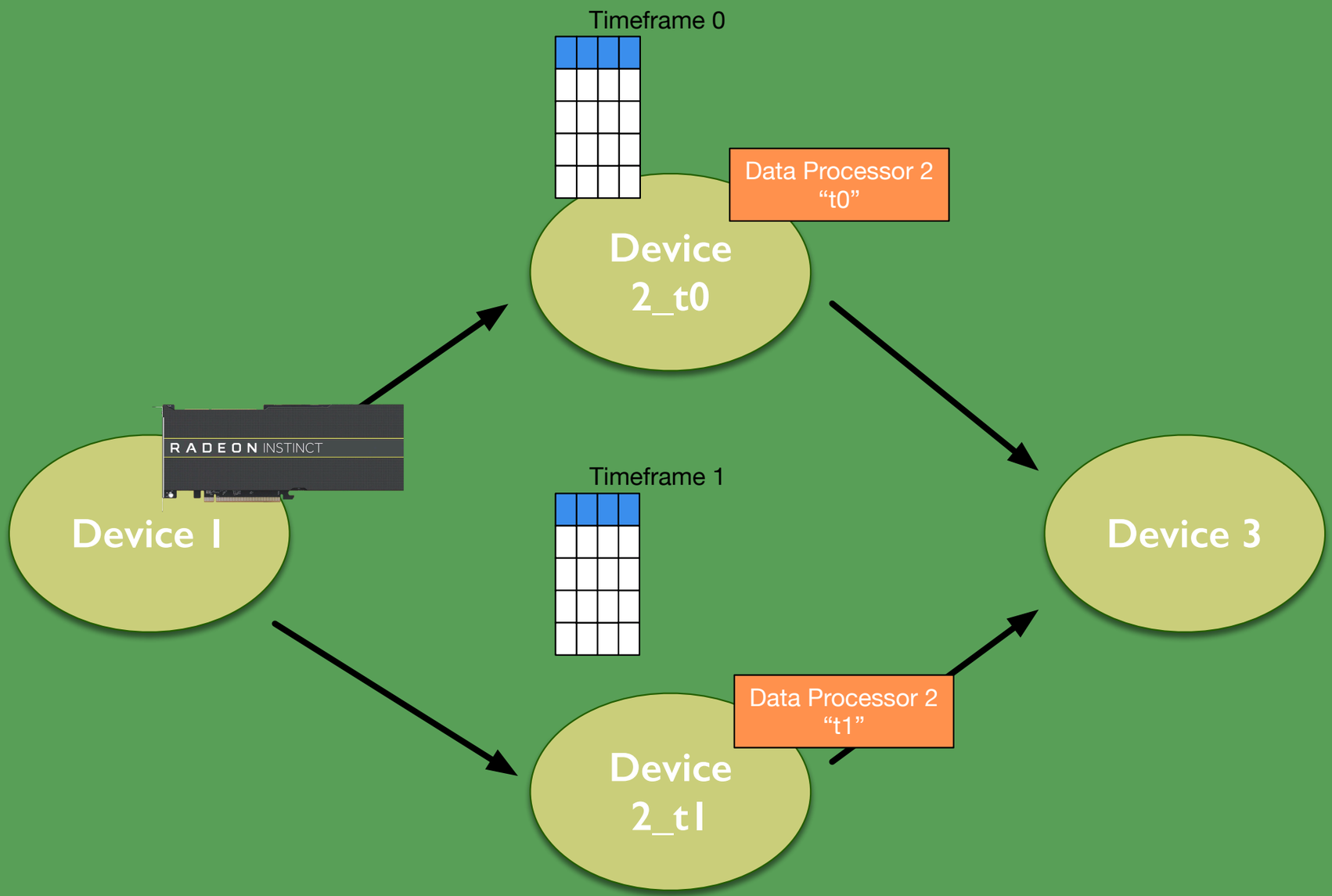
`(allocated - freed) < max-available-memory`



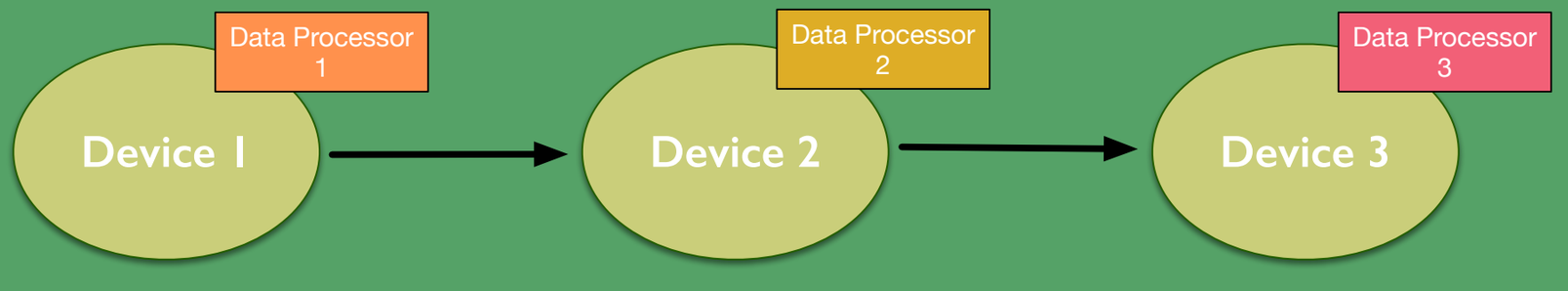
Works with memory as well...



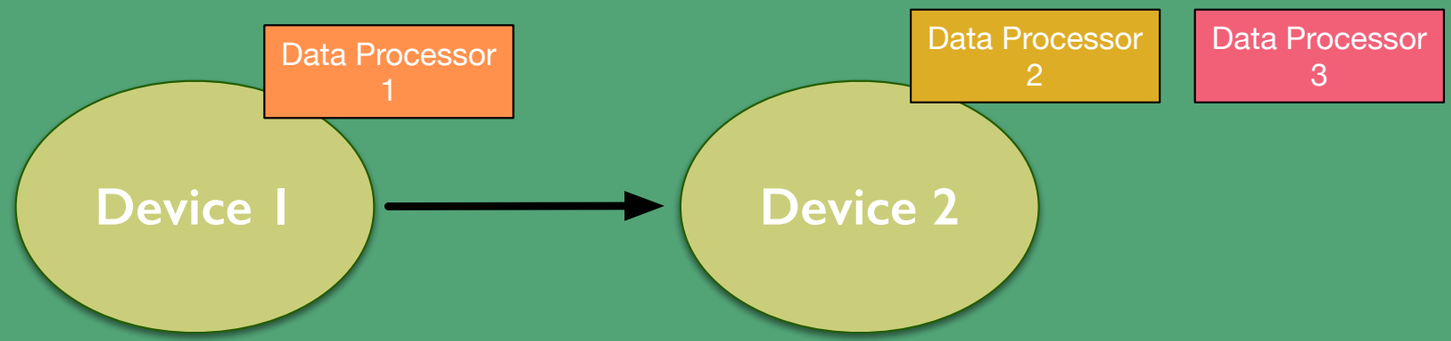
When GPUs are too fast...



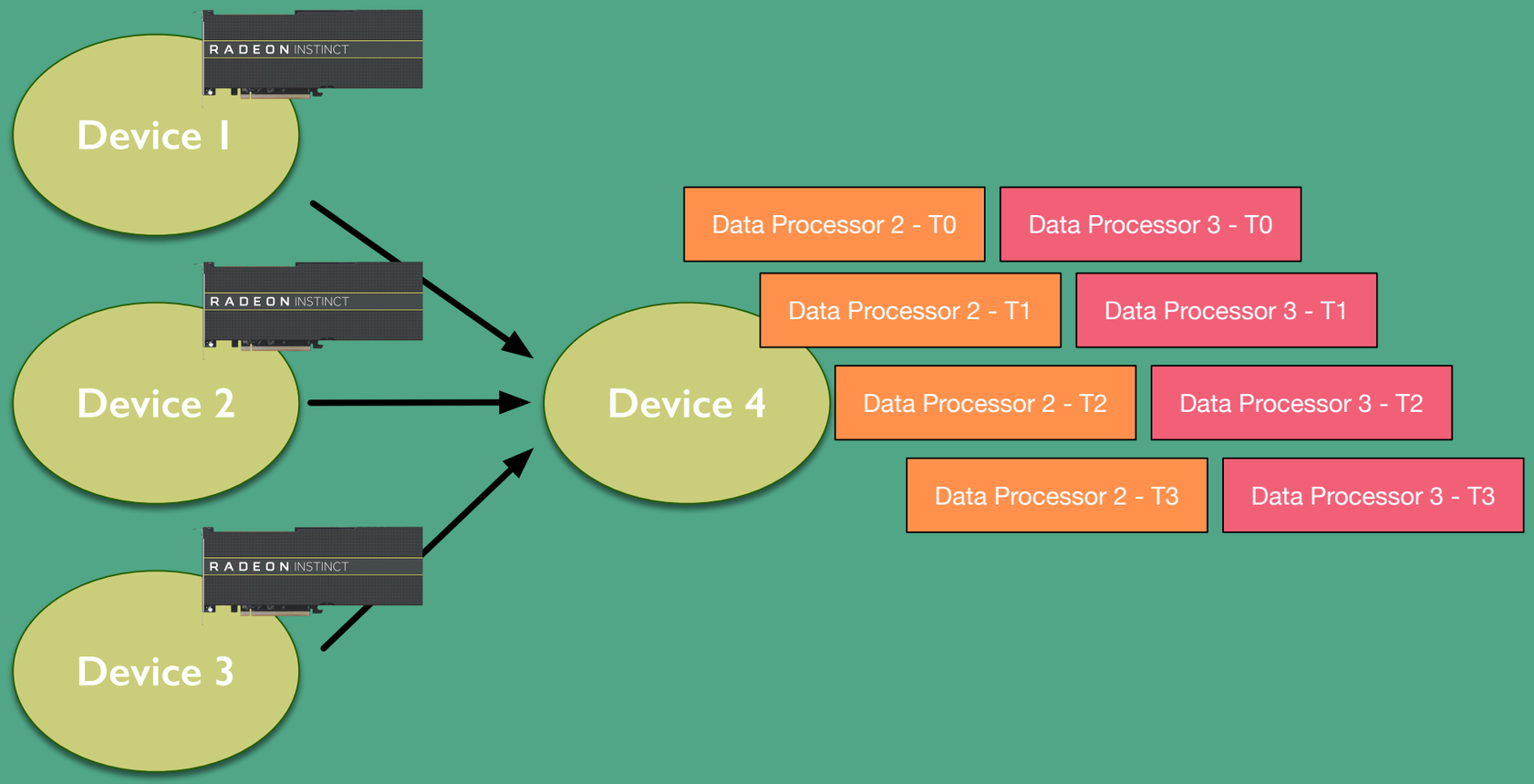
we can easily multiply downstream devices.



1-to-1 mapping Device <-> DPL Data Processor not mandatory!



Multiplexing



Seamless multithreading of data processors on a device is the ultimate goal