

NGT Task 3.2

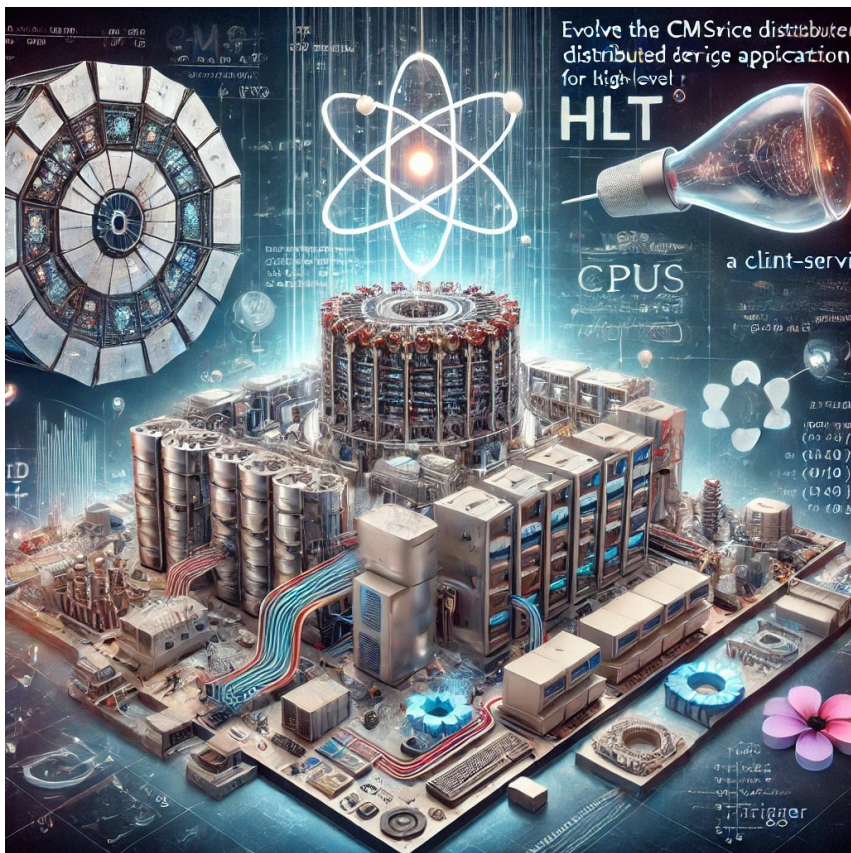
Evolving CMSSW into a client-server
distributed application for HLT

A. Bocci

EP/CMD



NextGen
Next Generation Triggers



the CMS Software (CMSSW)

- modular
 - overall more than 5000 different “modules”
 - e.g. HLT configuration for Run-3 composed of 4600 instances of 380 different modules
- parallel
 - multithreaded, good scalability over 100 of threads
- heterogeneous
 - alpaka-based modules
 - single source, built for CPUs, NVIDIA GPUs, AMD GPUs
 - transparent backend choice at runtime
- R&D: extend to multi-process/multi-node
 - single logical application, multiple machines
 - ALICE O2 is multi-process via message passing
 - ATLAS uses MPI in HPC environments

original use case

- GPU-equipped HLT farm
 - balance the amount of memory and processing power available on CPU and GPU
 - **fixed at time of procurement**
 - the HLT configuration and code base **evolves over the years**
- alternative approach
 - offload part of the GPU-heavy computations to separate nodes
 - increase GPU processing power over time simply adding more nodes
 - leverage high-speed network interconnect to minimise the extra latency

other use cases

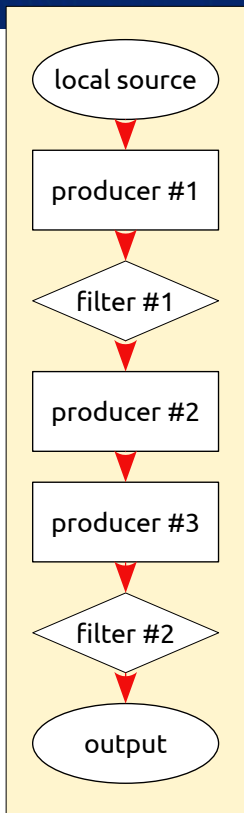
- ease deployment at HPC sites
 - gradual transition to GPU usage, mixing CPU-only nodes and GPU-heavy nodes
 - use worker nodes with limited local disk space or outbound network access
- mix and match jobs with different parallelism requirements
 - multi-threaded "client" communicating with multiple single-threaded "servers" or vice-versa

an ambitious goal

- **design and implement a distributed application**, with support for
 - sending and receiving arbitrary collections, including support inter-object references and provenance
 - multiple threads and multiple concurrent events (streams)
 - multiple senders and receivers per application
 - multiple clients and multiple servers with a non-trivial topology
 - efficient memory transfers to and from GPUs
 - event filtering
 - fault tolerance and error recovery
- **with minimal impact** on the existing code base
 - leverage the modular architecture of CMSSW and the Event Data Model approach
 - extend the framework capabilities
 - avoid rewriting and maintaining a dedicated reimplementaion of “remote” algorithms

simplified diagram

from

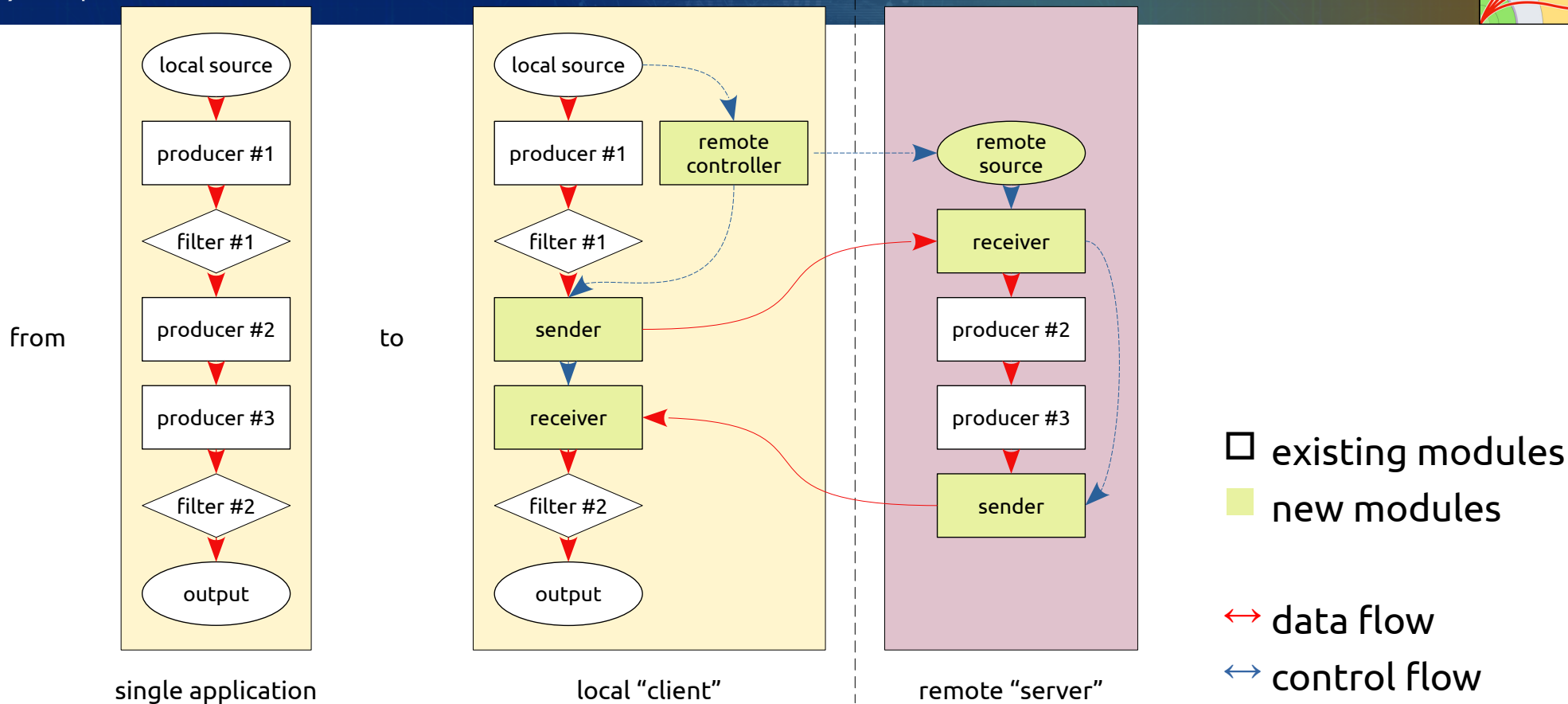


single application

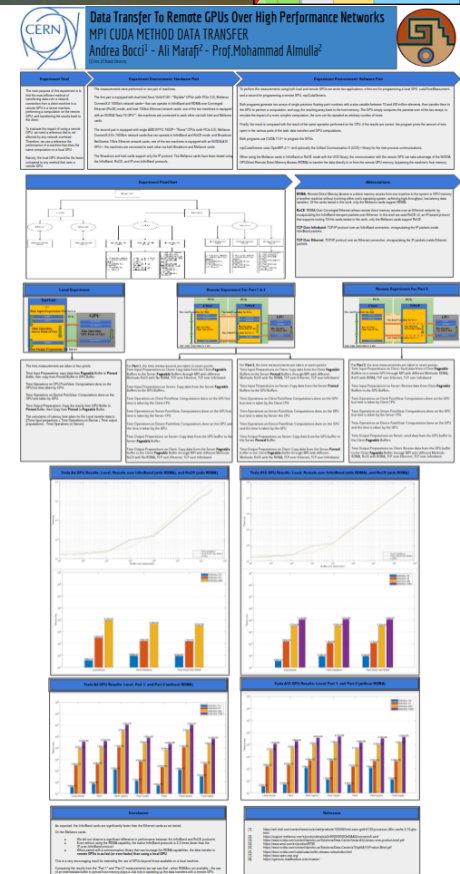
□ existing modules

↔ data flow

simplified diagram



- **modular approach** under study and evaluation since late 2020
 - limited personpower before NGT project
 - designed evolved during this year along with the first concrete implementation
 - thanks to Prof. Fawaz Alazemi (Kuwait University) and Andrea Valenzuela (CERN)
- choice to **use MPI for the prototype**
 - widely used in HPC (not so much in HEP)
 - efficient data movement
 - within a single node (shared memory) and over high speed interconnects (InfiniBand, RoCE, *etc.*)
 - support RDMA to and from GPU memory
 - demonstrated close-to-local data transfer performance (Ali Marafi, [ACAT 2022](#))
 - work on fault tolerance integrated in the **MPI 4.0 standard**
 - “User Level Fault Mitigation” available in [OpenMPI 5.0](#)
- validate these choices in the coming year(s)
 - measure the **performance** of the modules, library, and network solutions
 - evaluate **different paradigms** like RPC
 - Anna Polova will join the project as a technical student in February



Q4 2024

Implementation of a client-server, multithreaded, distributed test application, based on CMSSW, leveraging high-speed host-to-host or shared memory communication.

Q2 2025

Implementation of a small-scale demonstrator of a full HLT-like application.

Q2 2026

Support for optimal use of remote accelerators, *e.g.* using RDMA to/from GPU memory

Q4 2026

Support for multiple servers and distributed configurations.

Q4 2027

Compare different approaches to improve the resiliency of the system, such as server redundancy and client-side failure mitigation strategies.

Q2 2028

Evaluate the performance of different network interconnects and communication protocols.

Q4 2028

Large scale deployment and testing of the whole infrastructure in view of the HL-LHC data-taking in Run 4

✓ Q4 2024

Implementation of a client-server, multithreaded, distributed test application, based on CMSSW, leveraging high-speed host-to-host or shared memory communication.

Q2 2025

Implementation of a small-scale demonstrator of a full HLT-like application.

Q2 2026

Support for optimal use of remote accelerators, *e.g.* using RDMA to/from GPU memory

Q4 2026

Support for multiple servers and distributed configurations.

Q4 2027

Compare different approaches to improve the resiliency of the system, such as server redundancy and client-side failure mitigation strategies.

Q2 2028

Evaluate the performance of different network interconnects and communication protocols.

Q4 2028

Large scale deployment and testing of the whole infrastructure in view of the HL-LHC data-taking in Run 4



2024 results

☑ step 1

- controller/source
- no sender/receiver
- single thread, single stream
- single client, single server

☑ step 2

- send/receive fixed types
- single sender/receiver

☑ step 3

- multiple threads, multiple streams

☑ step 4

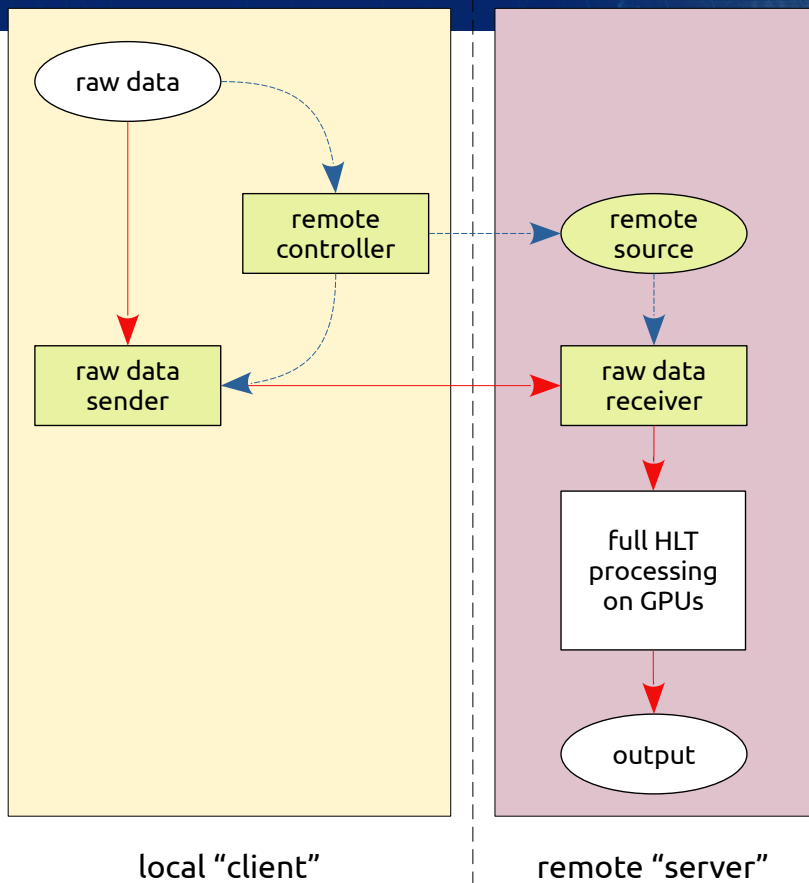
- multiple senders, multiple receivers

☑ 2024 demonstrator

- integrate steps 2, 3, 4
- controller/source
- send/receive fixed types
- multiple senders, multiple receivers
- multiple threads, multiple streams
- no support for `edm::Ref` and similar
- single client, single server

implemented in [cms-sw/cmssw#32632](https://github.com/cms-sw/cmssw/pull/32632)

fully remote processing



- **single-sided communication**
 - send RAW data from one process to another
 - full HLT reconstruction in the remote process
- **demonstrate controller / follower pattern**
 - establish communication
 - synchronise run, luminosity block and event transitions
- **demonstrate data distribution**
 - single collection type: RAW data
 - single sender
 - single receiver



HLT running over a simple MPI topology



```

$ CUDA_VISIBLE_DEVICES=0 numactl -N 2 cmsRun hit_local.py
%MSG-l ThreadStreamSetup: (NoModuleName) 20-Nov-2024 08:09:17 CET pre-events
setting # threads 32
setting # streams 24
%MSG

Trying to connect to the MPI server on port 2563244033.0:1645953384
Client connected to 1 server

%MSG-e EvFwdQDirector: AfterSourceConstruction 20-Nov-2024 08:09:22
problem with creating filedesc for furlwritelock - /gpu_data/store/data/Run2024F/EphemeralHLTPhysics/FED/run383631/fu.lock create:0
%MSG
%MSG-e EvFwdQDirector: AfterSourceConstruction 20-Nov-2024 08:09:22 CET pre-events
problem with opening furlwritelock file stream - Bad file descriptor
%MSG

Begin processing the 1st record. Run 383631, Event 982596563, LumiSec
Begin processing the 2nd record. Run 383631, Event 982924117, LumiSec
Begin processing the 3rd record. Run 383631, Event 981933835, LumiSec
Begin processing the 4th record. Run 383631, Event 981530982, LumiSec
Begin processing the 5th record. Run 383631, Event 982996637, LumiSec
Begin processing the 6th record. Run 383631, Event 981565582, LumiSec
Begin processing the 7th record. Run 383631, Event 981972131, LumiSec
Begin processing the 8th record. Run 383631, Event 982929497, LumiSec
Begin processing the 9th record. Run 383631, Event 981634919, LumiSec
Begin processing the 10th record. Run 383631, Event 981175354, LumiSec
Begin processing the 11th record. Run 383631, Event 983221457, LumiSec
Begin processing the 12th record. Run 383631, Event 983271661, LumiSec
Begin processing the 13th record. Run 383631, Event 983346129, LumiSec
Begin processing the 14th record. Run 383631, Event 981904883, LumiSec
Begin processing the 15th record. Run 383631, Event 983306223, LumiSec
Begin processing the 16th record. Run 383631, Event 982872151, LumiSec
Begin processing the 17th record. Run 383631, Event 982076029, LumiSec
Begin processing the 18th record. Run 383631, Event 981441688, LumiSec
Begin processing the 19th record. Run 383631, Event 983347938, LumiSec
Begin processing the 20th record. Run 383631, Event 981388786, LumiSec
Begin processing the 21st record. Run 383631, Event 983394385, LumiSec
Begin processing the 22nd record. Run 383631, Event 981166471, LumiSec
Begin processing the 23rd record. Run 383631, Event 982210937, LumiSec
Begin processing the 24th record. Run 383631, Event 982400609, LumiSec
Begin processing the 25th record. Run 383631, Event 982125473, LumiSec
Begin processing the 26th record. Run 383631, Event 982580418, LumiSec
Begin processing the 27th record. Run 383631, Event 983317911, LumiSec
Begin processing the 28th record. Run 383631, Event 983887255, LumiSec
Begin processing the 29th record. Run 383631, Event 982623268, LumiSec
Begin processing the 30th record. Run 383631, Event 983038883, LumiSec
Begin processing the 31st record. Run 383631, Event 981578598, LumiSec
Begin processing the 32nd record. Run 383631, Event 983390545, LumiSec
Begin processing the 33rd record. Run 383631, Event 982685289, LumiSec
Begin processing the 34th record. Run 383631, Event 981974171, LumiSec
Begin processing the 35th record. Run 383631, Event 983342768, LumiSec
Begin processing the 36th record. Run 383631, Event 983307378, LumiSec
Begin processing the 37th record. Run 383631, Event 981339841, LumiSec
Begin processing the 38th record. Run 383631, Event 981294226, LumiSec
Begin processing the 39th record. Run 383631, Event 981731744, LumiSec
Begin processing the 40th record. Run 383631, Event 982011720, LumiSec
Begin processing the 41st record. Run 383631, Event 981327077, LumiSec
Begin processing the 42nd record. Run 383631, Event 982401722, LumiSec
Begin processing the 43rd record. Run 383631, Event 980972448, LumiSec
Begin processing the 44th record. Run 383631, Event 981584972, LumiSec
Begin processing the 45th record. Run 383631, Event 981189246, LumiSec
Begin processing the 46th record. Run 383631, Event 982347017, LumiSec
Begin processing the 47th record. Run 383631, Event 982950317, LumiSec
Begin processing the 48th record. Run 383631, Event 981143888, LumiSec
Begin processing the 49th record. Run 383631, Event 983077378, LumiSec
Begin processing the 50th record. Run 383631, Event 981989628, LumiSec
Begin processing the 51st record. Run 383631, Event 981851098, LumiSec
Begin processing the 52nd record. Run 383631, Event 982235742, LumiSec
Begin processing the 53rd record. Run 383631, Event 982378714, LumiSec

```

local process without GPUs

MPI controller

DAQ source, reading local data

data are sent over MPI to the remote process

```

$ CUDA_VISIBLE_DEVICES=0 numactl -N 2 cmsRun hit_remote.py
%MSG-l ThreadStreamSetup: (NoModuleName) 20-Nov-2024 08:09:07 CET pre-events
setting # threads 32
setting # streams 24
%MSG

%MSG-i AlpaKaService: (NoModuleName) 20-Nov-2024 08:09:11 CET pre-events
AlpaKaServiceSerialSync successfully initialised.
Found 1 device:
- AMD EPYC 7763 64-Core Processor
%MSG

%MSG-i CUDAService: (NoModuleName) 20-Nov-2024 08:09:11 CET pre-events
CUDA runtime version 12.4, driver version 12.4, NVIDIA driver version 550.54.15
CUDA device 0: Tesla T4 (sm_75)
%MSG

%MSG-i AlpaKaService: (NoModuleName) 20-Nov-2024 08:09:11 CET pre-events
AlpaKaServiceCudaAsync successfully initialised.
Found 1 device:
- Tesla T4
%MSG

waiting for a connection to the MPI server at port 2563244033.0:1645953384
connected from 0

-----
#
# FastJet release 3.4.1
# M. Cacciari, G.P. Salam and G. Soyez
# A software package for jet finding and analysis at colliders
# http://fastjet.fr
#
# Please cite EPJCT2(2012)1896 [arXiv:1111.6097] if you use this package
# for scientific work and optionally PLB641(2006)57 [hep-ph/0512210].
#
# FastJet is provided without warranty under the GNU GPL v2 or higher.
# It uses T. Chan's closest pair algorithm, S. Fortune's Voronoi code
# and 3rd party plugin jet algorithms. See COPYING file for details.
#-----

%MSG-w EnergyInDeadEB FE: EcalRecHitProducer:hitEcalRecHit 20-Nov-2024 08:09:54 CET Run:
TP energy in the dead TT = 53 at (EcalTT subdet Barrel zc -ieta 17 iphi 37)
%MSG

%MSG-w BasicTrajectoryState: TrackProducer:hitL3NoFiltersTKTracksFromL2IOHitNoVtx 20-Nov-2024 08:09:57 CET Run: 383631 Event: 981546026
error not pos-def:
[ 2.35041e+07 5355.37 -2.21700e+06 -1458.34 4117.57
5355.37 1.22164 -505.157 -0.331914 0.961726
-2.21700e+06 -505.157 209130 137.456 -388.397
-1458.34 -0.331914 137.456 -0.786224 -0.237483
4117.57 0.961726 -388.397 -0.237483 1.10983 ]
pos/mom/mf (11.0467,-1.83103,11.7456) (2.99099,-0.386183,5.14116) (0,0,3.81094)
%MSG

%MSG-w BasicTrajectoryState: TrackProducer:hitL3NoFiltersTKTracksFromL2IOHitNoVtx 20-Nov-2024 08:09:57 CET Run: 383631 Event: 981546026
local error not pos-def:
[ 2.35041e+07 2.2357e+06 310184 1462.79 -7927.32
2.2357e+06 212658 29504.5 139.033 -754.051
310184 29504.5 4893.51 19.2901 -104.435
1462.79 139.033 19.2901 -0.792368 -0.588296
-7927.32 -754.051 -104.435 -0.588296 4.18471 ]
pos/mom/mf (11.0467,-1.83103,11.7456) (2.99099,-0.386183,5.14116) (0,0,3.81094)
%MSG

%MSG-w BasicTrajectoryState: TrackProducer:hitL3NoFiltersTKTracksFromL2IOHitNoVtx 20-Nov-2024 08:09:57 CET Run: 383631 Event: 981546026
error not pos-def:
[ 90607.5 -336.094 -49194.8 -646496 22183.5
-336.094 -1.42157 184.21 2833.51 -97.2274
-49194.8 184.21 28684.7 353236 -121119.9
-646496 2833.51 353236 4.58142e+06 -157211
22183.5 -97.2274 -121119.9 -157211 4004.06 ]
pos/mom/mf (33.2229,-1.83651,48.5723) (1.97818,0.0892046,3.03378) (0,0,3.80245)
%MSG

```

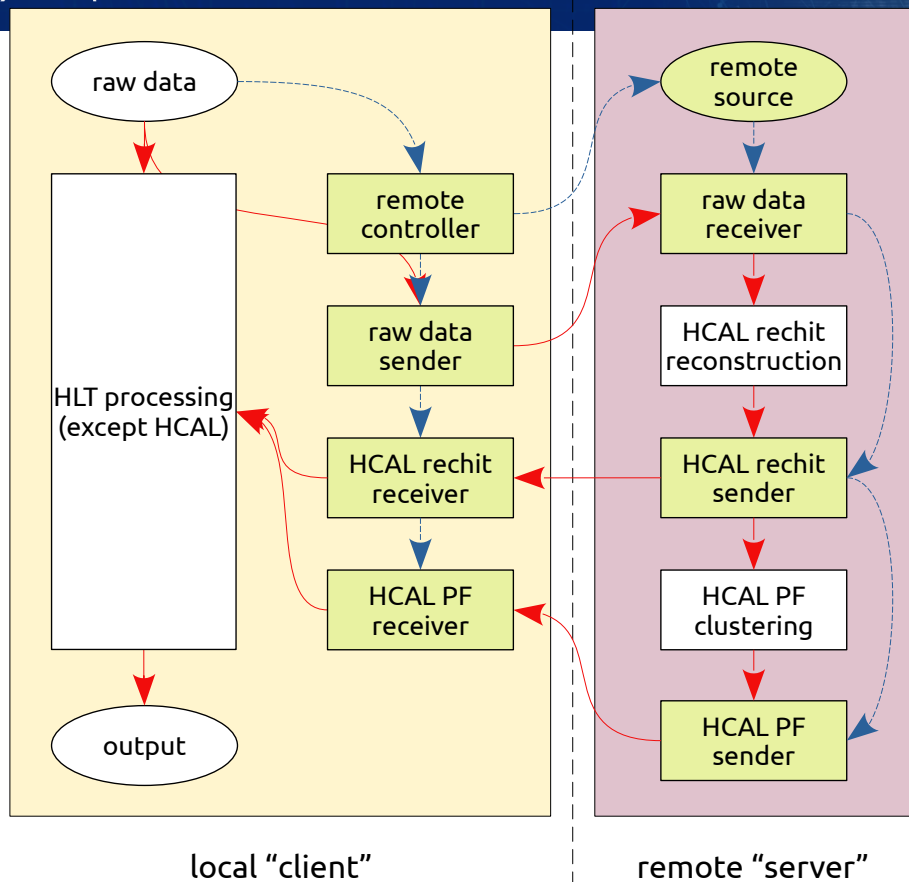
remote process with one GPU

MPI source

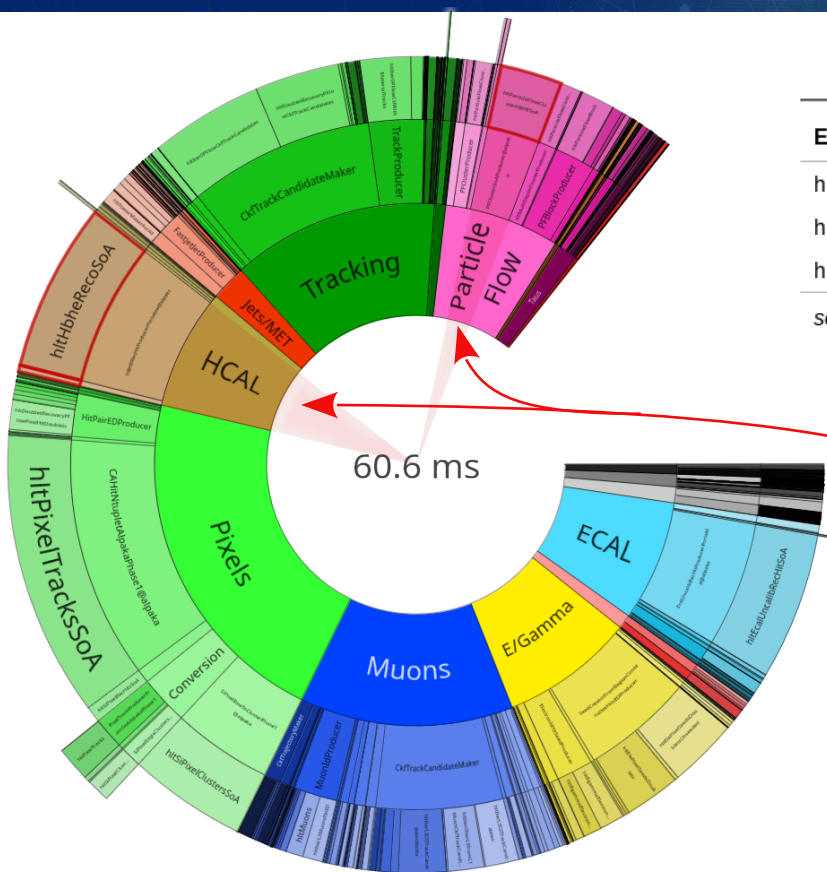
data are received over MPI, processed, filtered and stored



prototype of distributed processing



- multi-sided communication
 - send RAW data from one process to another
 - send back reconstructed objects
- demonstrate cooperative processing
 - run locally using only the CPU
 - offload part of processing to remote GPU node
- demonstrate data distribution
 - multiple collection types
 - RAW data
 - HCAL rechit SoA
 - HCAL PF cluster SoA
- multiple senders, multiple receivers



Element	Time	Fraction
hltHbheRecoSoA	43.2 ms	6.7 %
hltHcalDigis	1.6 ms	0.2 %
hltParticleFlowClusterHBHESoA	15.8 ms	2.4 %
<i>selected</i>	60.6 ms	9.4 %

- ~90% of the HLT runs locally on CPU
- ~10% runs remotely on GPU





a distributed HLT application over MPI



```

$ CUDA_VISIBLE_DEVICES= numactl -N 0 cmsRun hlt_local.py
%MSG-i ThreadStreamSetup: (NoModuleName) 27-Nov-2024 01:28:07 CET
setting # threads 32
setting # streams 24
%MSG
%MSG-i AlpakaService: (NoModuleName) 27-Nov-2024 01:28:09 CET pre-events
AlpakaServiceSerialSync successfully initialised.
Found 1 device:
  - AMD EPYC 9754 128-Core Processor
%MSG

Trying to connect to the MPI server on port 2770534401.0:611459426
Client connected to 1 server
receive type "PortableHostCollection-hcal:HcalRechtIsoLayout<128,false>" for label "" over MPI channel instance 11
receive type "PortableHostCollection-reco:PFRechtIsoLayout<128,false>" for label "" over MPI channel instance 12
send branch "FEDRawDataCollection_rawDataCollector_LHC" of type "FEDRawDataCollection" over MPI channel instance 1
%MSG
%MSG-e EvFDAQDirector: AfterSourceConstruction 27-Nov-2024 01:28:42 CET pre-events
problem with creating filedesc for fwurltsock -- /gpu_data/store/data/Run2024F/EphemeralHLTPhysics/FED/Run383631/fu.lock
%MSG
%MSG-e EvFDAQDirector: AfterSourceConstruction 27-Nov-2024 01:28:
problem with opening fwurltsock file stream -- Bad file descriptor
%MSG

-----
# FastJet release 3.4.1
# M. Cacciari, G.P. Salam and G. Soyez
# A software package for jet finding and analysis at colliders
# http://fastjet.fr
#
# Please cite EPJ C72(2012)1896 [arXiv:1111.6097] if you use this package
# For scientific work and optionally PLB641(2006)157 [hep-ph/0512210].
#
# FastJet is provided without warranty under the GNU GPL v2 or higher
# It uses T. Chan's closest pair algorithm, S. Fortnes's Voronoi
# and 3rd party plugin jet algorithms. See COPYING file for details
#-----

Begin processing the 1st record. Run 383631, Event 982586563, LumiSection 410 on stream 0 at 27-Nov-2024 01:28:45.957 CET
Begin processing the 2nd record. Run 383631, Event 982924117, LumiSection 410 on stream 22 at 27-Nov-2024 01:28:45.958 CET
Begin processing the 3rd record. Run 383631, Event 981933835, LumiSection 410 on stream 23 at 27-Nov-2024 01:28:45.958 CET
Begin processing the 4th record. Run 383631, Event 981530982, LumiSection 410 on stream 23 at 27-Nov-2024 01:28:45.958 CET
Begin processing the 5th record. Run 383631, Event 982090637, LumiSection 410 on stream 23 at 27-Nov-2024 01:28:45.958 CET
Begin processing the 6th record. Run 383631, Event 981565582, LumiSection 410 on stream 23 at 27-Nov-2024 01:28:45.960 CET
Begin processing the 7th record. Run 383631, Event 981972131, LumiSection 410 on stream 21 at 27-Nov-2024 01:28:45.961 CET
Begin processing the 8th record. Run 383631, Event 982929497, LumiSection 410 on stream 7 at 27-Nov-2024 01:28:45.963 CET
Begin processing the 9th record. Run 383631, Event 981904883, LumiSection 410 on stream 16 at 27-Nov-2024 01:28:45.972 CET
Begin processing the 10th record. Run 383631, Event 981629419, LumiSection 410 on stream 6 at 27-Nov-2024 01:28:45.966 CET
Begin processing the 11th record. Run 383631, Event 983221457, LumiSection 410 on stream 20 at 27-Nov-2024 01:28:45.966 CET
Begin processing the 12th record. Run 383631, Event 983271661, LumiSection 410 on stream 2 at 27-Nov-2024 01:28:45.969 CET
Begin processing the 13th record. Run 383631, Event 983346129, LumiSection 410 on stream 13 at 27-Nov-2024 01:28:45.971 CET
Begin processing the 14th record. Run 383631, Event 981441688, LumiSection 410 on stream 15 at 27-Nov-2024 01:28:45.972 CET
Begin processing the 15th record. Run 383631, Event 983306223, LumiSection 410 on stream 3 at 27-Nov-2024 01:28:45.974 CET
Begin processing the 16th record. Run 383631, Event 982872151, LumiSection 410 on stream 0 at 27-Nov-2024 01:28:45.976 CET
Begin processing the 17th record. Run 383631, Event 982076029, LumiSection 410 on stream 19 at 27-Nov-2024 01:28:45.978 CET
Begin processing the 18th record. Run 383631, Event 981441688, LumiSection 410 on stream 18 at 27-Nov-2024 01:28:45.979 CET
Begin processing the 19th record. Run 383631, Event 983347930, LumiSection 410 on stream 15 at 27-Nov-2024 01:28:45.981 CET
Begin processing the 20th record. Run 383631, Event 981388786, LumiSection 410 on stream 8 at 27-Nov-2024 01:28:45.983 CET
Begin processing the 21st record. Run 383631, Event 983394385, LumiSection 410 on stream 10 at 27-Nov-2024 01:28:45.985 CET
Begin processing the 22nd record. Run 383631, Event 981166471, LumiSection 410 on stream 22 at 27-Nov-2024 01:28:45.986 CET
Begin processing the 23rd record. Run 383631, Event 982210937, LumiSection 410 on stream 17 at 27-Nov-2024 01:28:45.988 CET
Begin processing the 24th record. Run 383631, Event 982406089, LumiSection 410 on stream 4 at 27-Nov-2024 01:28:45.989 CET
Begin processing the 25th record. Run 383631, Event 982124473, LumiSection 410 on stream 1 at 27-Nov-2024 01:28:51.991 CET
Begin processing the 26th record. Run 383631, Event 982508418, LumiSection 410 on stream 7 at 27-Nov-2024 01:28:51.975 CET
Begin processing the 27th record. Run 383631, Event 981463791, LumiSection 410 on stream 23 at 27-Nov-2024 01:28:51.976 CET
Begin processing the 28th record. Run 383631, Event 983087255, LumiSection 410 on stream 10 at 27-Nov-2024 01:28:51.994 CET
Begin processing the 29th record. Run 383631, Event 982622269, LumiSection 410 on stream 8 at 27-Nov-2024 01:28:51.998 CET
Begin processing the 30th record. Run 383631, Event 983630883, LumiSection 410 on stream 0 at 27-Nov-2024 01:28:52.001 CET
Begin processing the 31st record. Run 383631, Event 981578590, LumiSection 410 on stream 14 at 27-Nov-2024 01:28:52.002 CET
Begin processing the 32nd record. Run 383631, Event 983300545, LumiSection 410 on stream 4 at 27-Nov-2024 01:28:52.007 CET

```

local process without GPUs

MPI modules

DAQ source, reading local data

local HLT processing, except HCAL part

```

$ CUDA_VISIBLE_DEVICES=0 numactl -N 3 cmsRun hlt_remote.py
%MSG-i ThreadStreamSetup: (NoModuleName) 27-Nov-2024 01:27:57 CET pre-events
setting # threads 32
setting # streams 24
%MSG
%MSG-i AlpakaService: (NoModuleName) 27-Nov-2024 01:27:58 CET pre-events
AlpakaServiceSerialSync successfully initialised.
Found 1 device:
  - AMD EPYC 9754 128-Core Processor
%MSG
%MSG-i CUDAService: (NoModuleName) 27-Nov-2024 01:27:58 CET pre-events
CUDA runtime version 12.4, driver version 12.4, NVIDIA driver version 550.54.15
CUDA device 0: NVIDIA L4 (sm_89)
%MSG
%MSG-i AlpakaService: (NoModuleName) 27-Nov-2024 01:27:58 CET pre-events
AlpakaServiceCudaAsync successfully initialised.
Found 1 device:
  - NVIDIA L4
%MSG

waiting for a connection to the MPI server at port 2770534401.0:611459426
receive type "FEDRawDataCollection" for label "" over MPI channel instance 1
send branch "128falsehcalHcalRechtIsoLayoutPortableHostCollection_HtHbHeRecoSOA_REM" of type "PortableHostCollection-hcal:HcalRechtIsoLayout<128,false>" over MPI channel instance 11
send branch "128falsehcalHcalRechtIsoLayoutPortableHostCollection_HtParticleFlowRechtIsoLayout" of type "PortableHostCollection-reco:PFRechtIsoLayout<128,false>" over MPI channel instance 12
connected from 0
new ProcessHistory registered:
Process History
  LHC 9168f19d9f0566b9d08b7495d6a1c847 "CMSWS_14_2_0_pre3" ;
  HLT 3ab9e9c7dccc7454143cfff00188787ed "CMSWS_14_2_0_pre3" ;

Begin processing the 1st record. Run 383631, Event 982586563, LumiSection 410 on stream 0 at 27-Nov-2024 01:28:45.957 CET
Begin processing the 2nd record. Run 383631, Event 982924117, LumiSection 410 on stream 22 at 27-Nov-2024 01:28:45.958 CET
Begin processing the 3rd record. Run 383631, Event 981933835, LumiSection 410 on stream 23 at 27-Nov-2024 01:28:45.958 CET
Begin processing the 4th record. Run 383631, Event 981530982, LumiSection 410 on stream 23 at 27-Nov-2024 01:28:45.958 CET
Begin processing the 5th record. Run 383631, Event 982090637, LumiSection 410 on stream 23 at 27-Nov-2024 01:28:45.958 CET
Begin processing the 6th record. Run 383631, Event 981565582, LumiSection 410 on stream 23 at 27-Nov-2024 01:28:45.960 CET
Begin processing the 7th record. Run 383631, Event 981972131, LumiSection 410 on stream 21 at 27-Nov-2024 01:28:45.961 CET
Begin processing the 8th record. Run 383631, Event 982929497, LumiSection 410 on stream 7 at 27-Nov-2024 01:28:45.963 CET
Begin processing the 9th record. Run 383631, Event 981904883, LumiSection 410 on stream 16 at 27-Nov-2024 01:28:45.972 CET
Begin processing the 10th record. Run 383631, Event 981629419, LumiSection 410 on stream 6 at 27-Nov-2024 01:28:45.966 CET
Begin processing the 11th record. Run 383631, Event 983221457, LumiSection 410 on stream 20 at 27-Nov-2024 01:28:45.966 CET
Begin processing the 12th record. Run 383631, Event 983271661, LumiSection 410 on stream 2 at 27-Nov-2024 01:28:45.969 CET
Begin processing the 13th record. Run 383631, Event 983346129, LumiSection 410 on stream 13 at 27-Nov-2024 01:28:45.971 CET
Begin processing the 14th record. Run 383631, Event 981441688, LumiSection 410 on stream 15 at 27-Nov-2024 01:28:45.972 CET
Begin processing the 15th record. Run 383631, Event 983306223, LumiSection 410 on stream 3 at 27-Nov-2024 01:28:45.974 CET
Begin processing the 16th record. Run 383631, Event 982872151, LumiSection 410 on stream 0 at 27-Nov-2024 01:28:45.976 CET
Begin processing the 17th record. Run 383631, Event 982076029, LumiSection 410 on stream 19 at 27-Nov-2024 01:28:45.978 CET
Begin processing the 18th record. Run 383631, Event 982076029, LumiSection 410 on stream 18 at 27-Nov-2024 01:28:45.979 CET
Begin processing the 19th record. Run 383631, Event 982210937, LumiSection 410 on stream 17 at 27-Nov-2024 01:28:45.981 CET
Begin processing the 20th record. Run 383631, Event 981388786, LumiSection 410 on stream 8 at 27-Nov-2024 01:28:45.983 CET
Begin processing the 21st record. Run 383631, Event 983394385, LumiSection 410 on stream 10 at 27-Nov-2024 01:28:45.985 CET
Begin processing the 22nd record. Run 383631, Event 981166471, LumiSection 410 on stream 22 at 27-Nov-2024 01:28:45.986 CET
Begin processing the 23rd record. Run 383631, Event 982210937, LumiSection 410 on stream 17 at 27-Nov-2024 01:28:45.988 CET
Begin processing the 24th record. Run 383631, Event 982406089, LumiSection 410 on stream 4 at 27-Nov-2024 01:28:45.989 CET
Begin processing the 25th record. Run 383631, Event 983630883, LumiSection 410 on stream 1 at 27-Nov-2024 01:28:51.991 CET
Begin processing the 26th record. Run 383631, Event 982508418, LumiSection 410 on stream 7 at 27-Nov-2024 01:28:51.975 CET
Begin processing the 27th record. Run 383631, Event 981463791, LumiSection 410 on stream 23 at 27-Nov-2024 01:28:51.976 CET
Begin processing the 28th record. Run 383631, Event 983087255, LumiSection 410 on stream 10 at 27-Nov-2024 01:28:51.994 CET
Begin processing the 29th record. Run 383631, Event 982622269, LumiSection 410 on stream 8 at 27-Nov-2024 01:28:51.998 CET
Begin processing the 30th record. Run 383631, Event 983630883, LumiSection 410 on stream 0 at 27-Nov-2024 01:28:52.001 CET
Begin processing the 31st record. Run 383631, Event 981578590, LumiSection 410 on stream 14 at 27-Nov-2024 01:28:52.002 CET
Begin processing the 32nd record. Run 383631, Event 983300545, LumiSection 410 on stream 4 at 27-Nov-2024 01:28:52.007 CET

```

remote process with one GPU

MPI modules

raw data are received over MPI

HCAL reconstructed collections are sent back over MPI





the road ahead

already in
progress !

- 2025 deliverables
 - simplify the code base and support send/receive of arbitrary collections
 - efficient encoding and decoding of "Structure of Arrays" data types
 - bypass ROOT de/serialisation for types with a known layout
 - in collaboration with Task 3.1.2 and Task 1.7
- 2026 deliverables
 - send data directly from (local) GPU memory to (remote) GPU memory
 - non-linear topology with multiple clients and servers
- 2027 deliverables and **contractual milestone**
 - study various fault tolerance approaches
 - **implementation of a client-server, multithreaded, distributed test application, based on the CMS software framework CMSSW, leveraging high-speed host-to-host or shared memory communication**
- 2028 deliverables
 - study different hardware interconnects and software libraries
 - plan for large-scale deployment in Run 4

long-term goal,
start next year





NextGen
Next Generation Triggers



NexTGen
Next Generation Triggers