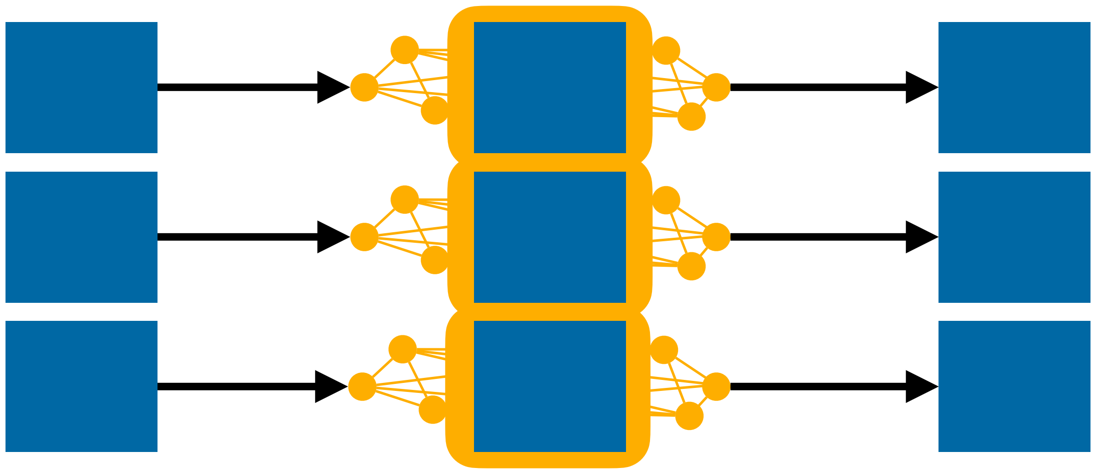
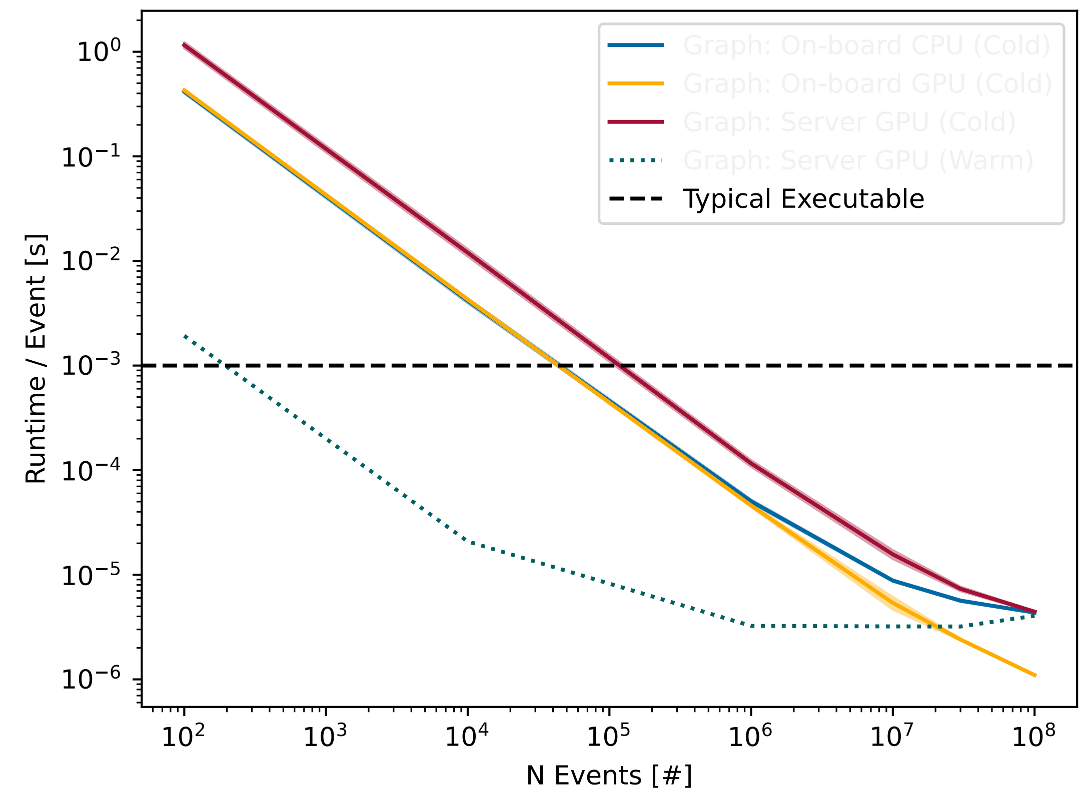




Vectorised Neutrino Reconstruction by Computing Graphs

Martin Erdmann, Peter Fackeldey, Benjamin Fischer, **Dennis Noll**

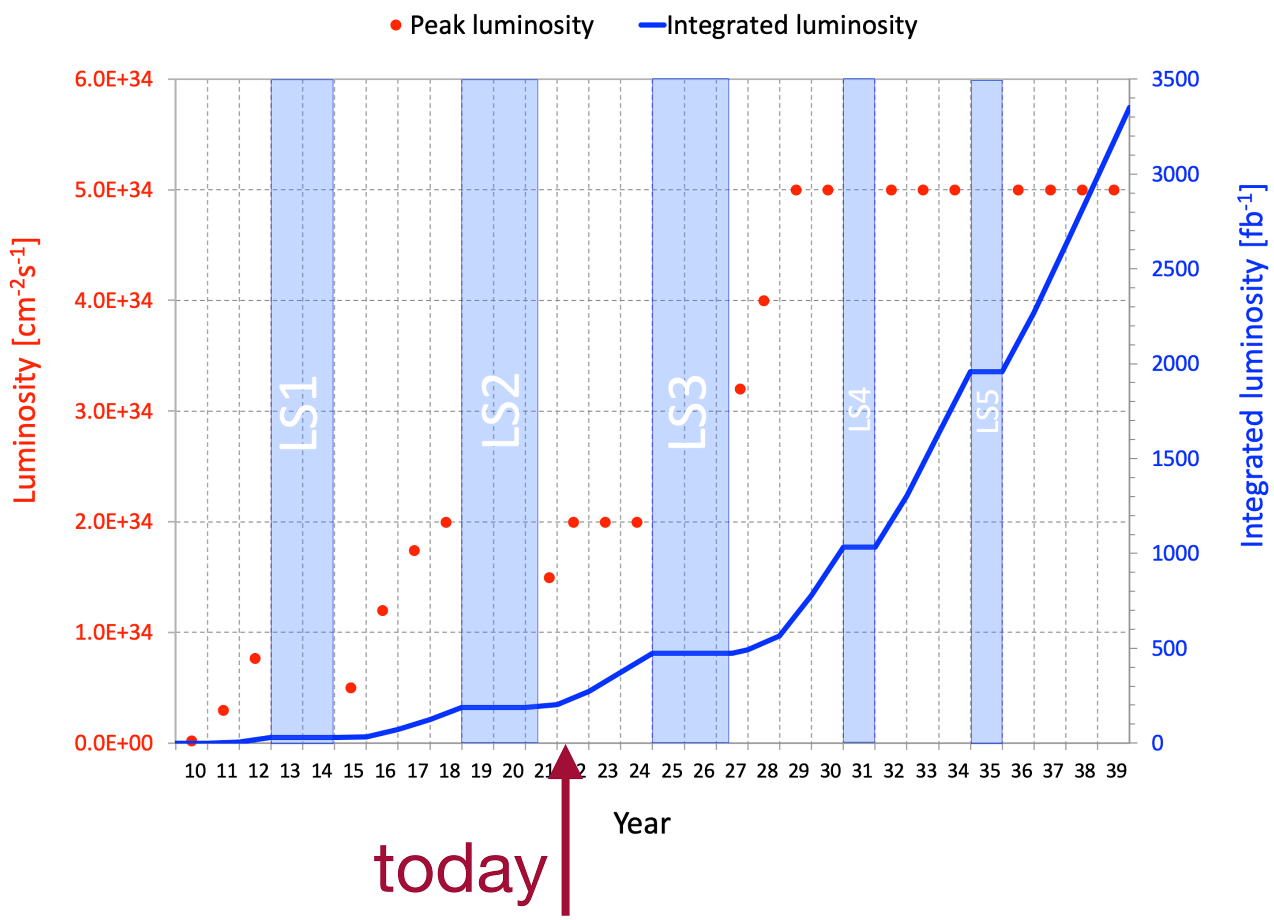
IDT-UM Collaboration Meeting

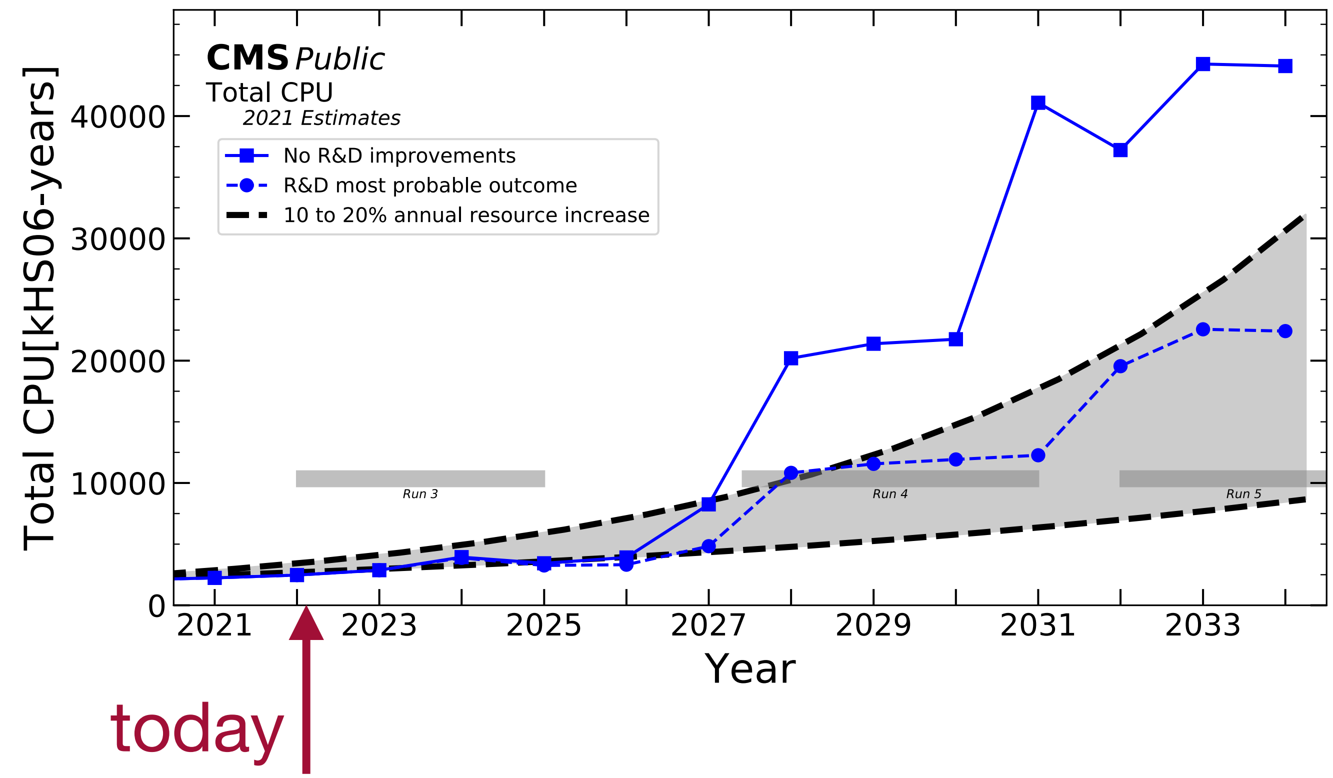
14.02.22

RWTH AACHEN UNIVERSITY

RWTH AACHEN UNIVERSITY
Dennis Noll - 14.02.22

Motivation





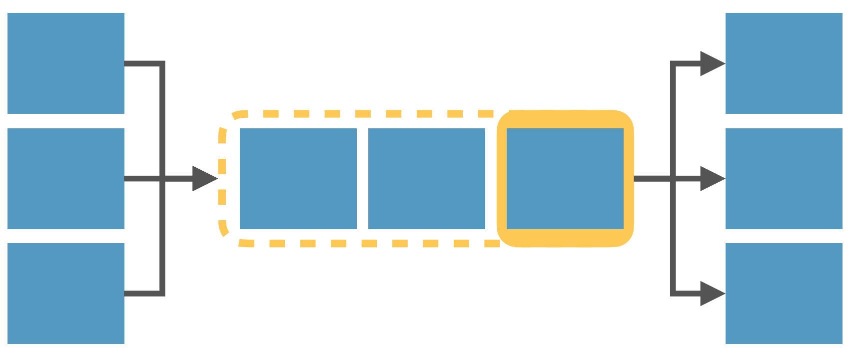
- Fast turn-around times of HEP analyses are driver of scientific insight
- Traditional analyses already O(weeks)
- Data increased in HL-LHC by x20
- Future analyse must be: Faster & More resource efficient

➜ Requires re-thinking of analysis computing!

Vectorised Computations

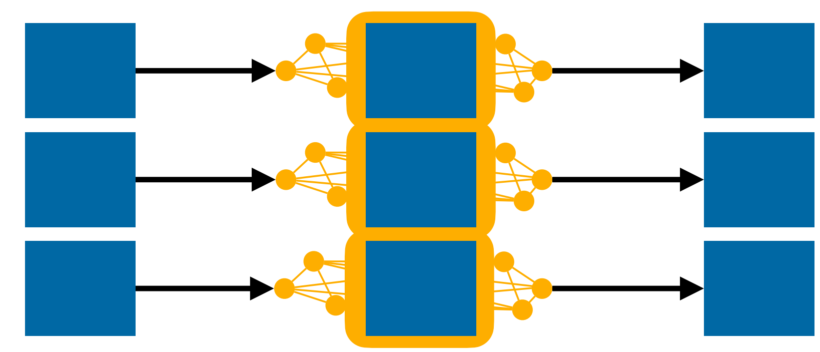
- Fast and efficient analyses can be realised with vectorised computations
- Already used by many analyses

Event loop



- Load one event
- Evaluate expressions
- Store results
- Repeat

Vectorised

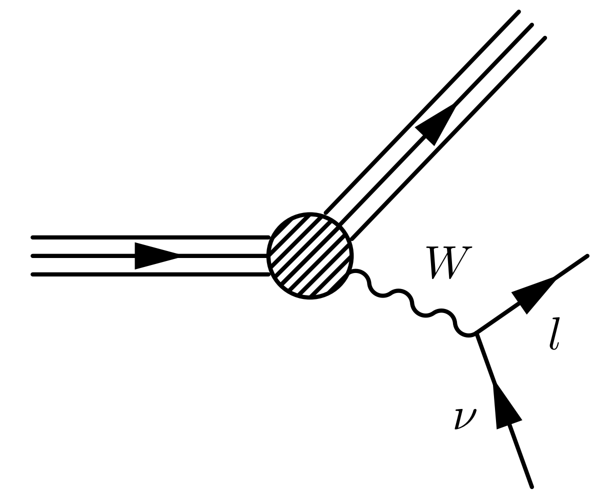


- Load **many** events
- Evaluate **vectorised** expressions
- Store results
- Repeat

- Problem: Some computations challenging to vectorise!
- E.g. Neutrino reconstruction (next page)

Physics Case: Neutrino Reconstruction

- Reconstruction of longitudinal neutrino momentum in e.g. ttH events
- Solved by assuming: $\not{E}_T = p_{\nu,T}$, W mass
- Inputs: Lepton, $\not{E}_T$
- Two branches:
  - Real branch (h≦1): Purely analytical
  - Complex branch (h>1): Involves fitting



$$p_{\nu,z}^{1,2} = \frac{k}{p_{l,T}^2}\left(p_{l,z} \pm E_l \sqrt{1 - \underbrace{\left(\frac{p_{l,T}\, p_{\nu,T}}{k}\right)^2}_{\equiv h}}\right)$$

Optimal benchmark

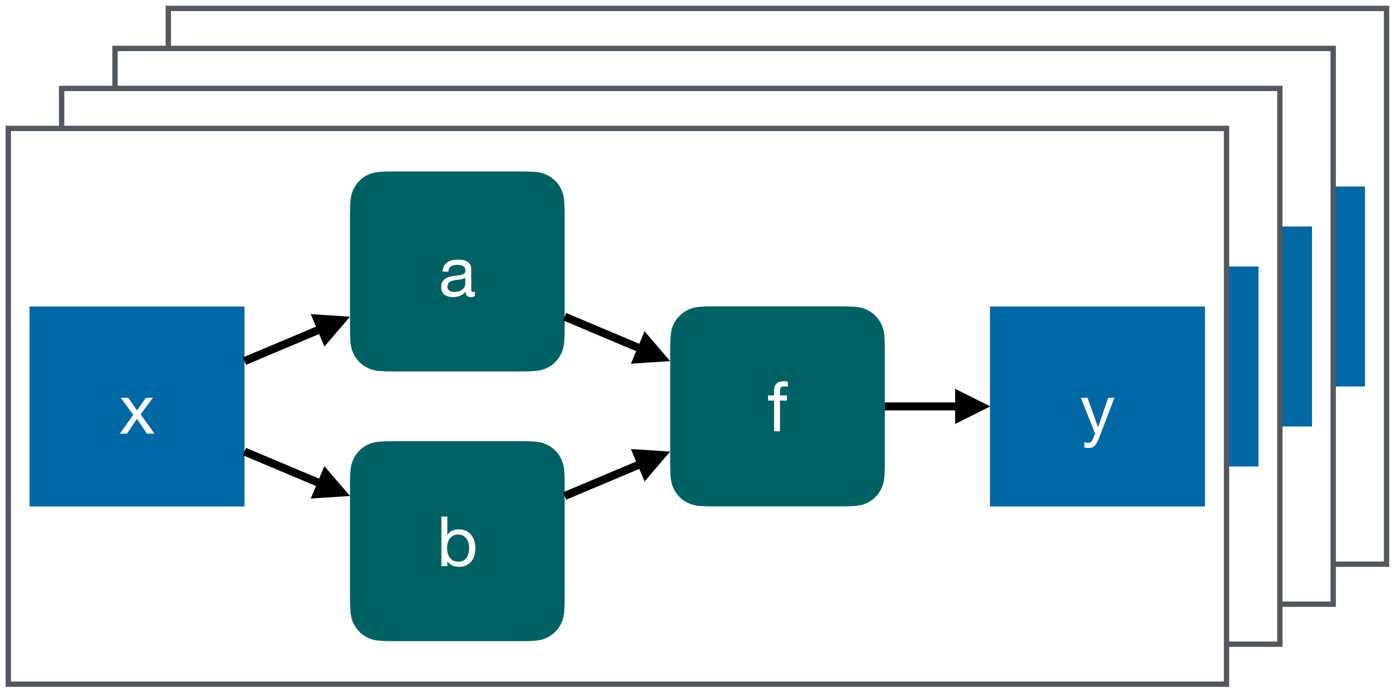
- Multiple Inputs (Lepton, MET)
- Different behaviour on event-basis
- Stateful computation: Fitting
- Physics result can easily be verified

➜ Vectorise using graph computing model!

Computing Graph

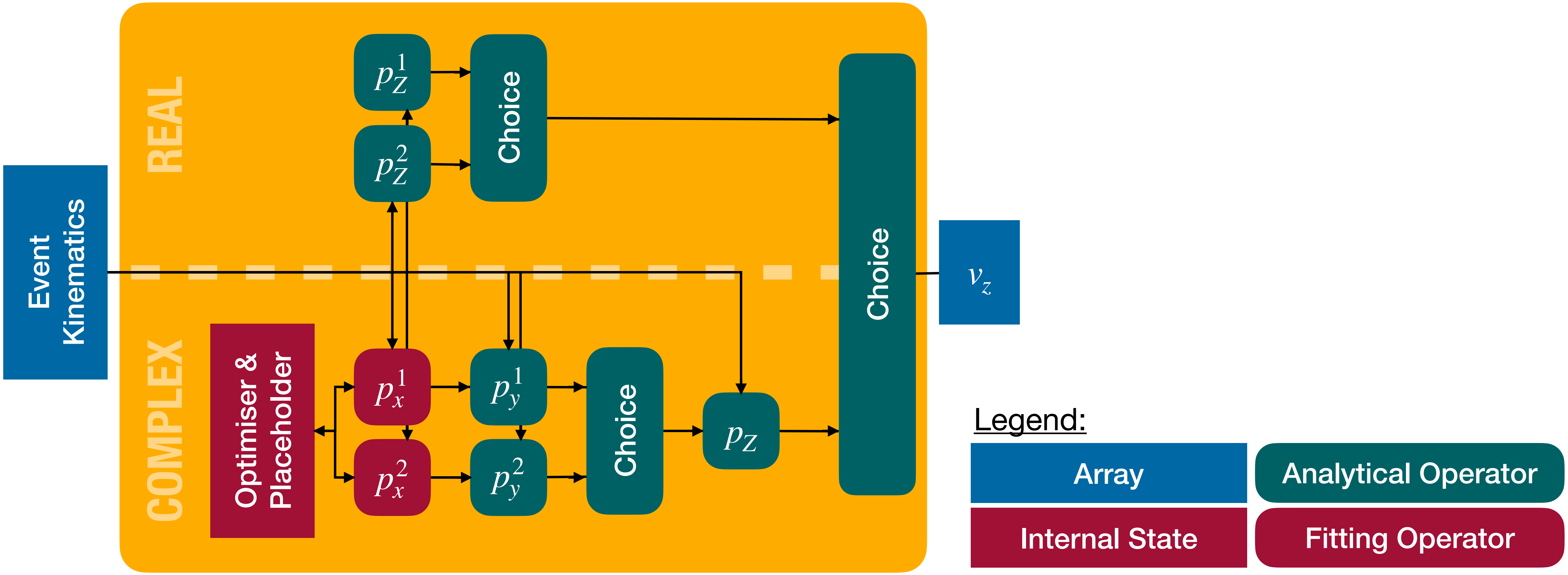
- Contents:
  - Nodes = Computations
  - Edges = Data flow
- Properties:
  - Directed = ⟶
  - Acyclic = no loops

Example y = f(a(x), b(x)):



- Two levels of parallelism:
  - Inter processing unit:
    - Parallel units in directed acyclic graph (e.g. a & b) can run in parallel
  - Intra processing unit (SIMD):
    - If graph is same for multiple inputs (N_{events})
    - Parallel execution over many events

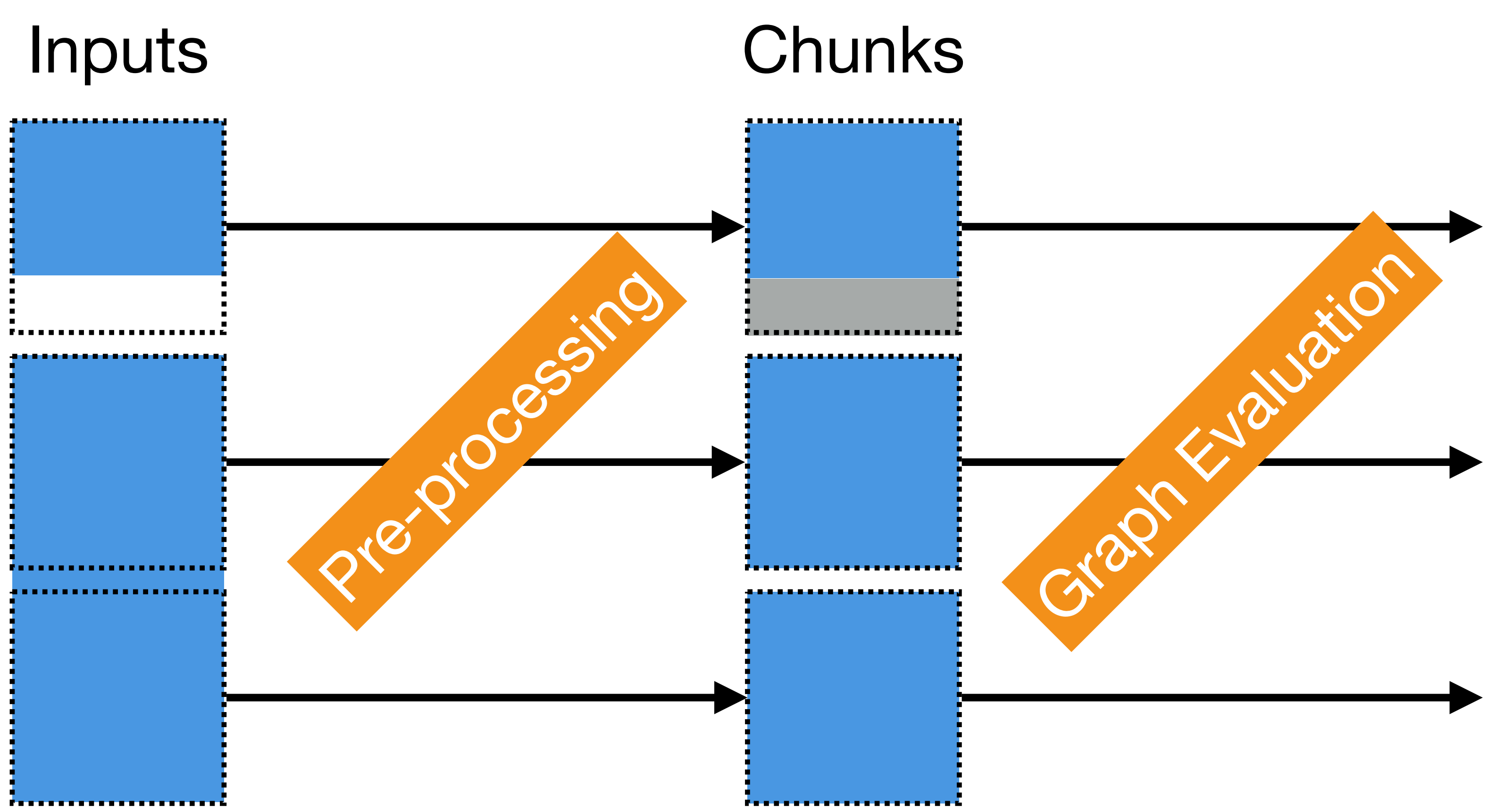
Neutrino Reconstruction in a Graph



- Two branches (real and complex)
- Fit of neutrino momentum:
    - Unrolled for-loop with 100 iterations
    - Using ADAM optimiser
- Conditions (choices) guide logical rather than physical flow
    - All expressions evaluated
    - Graph is the same for every event

Implementation of Computing Graph

- Graph implemented with TensorFlow:
  - Supports processing on GPU
  - Wrapped in Keras model:
    - Portable (saving to/loading from disk)
    - Straight forward integration
- Pre-processing:
  - Structure of graph must be static
  - Requires batching of events in fixed size chunks



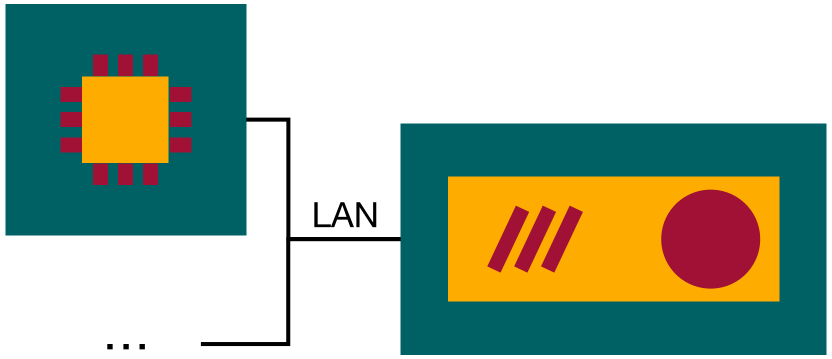
Experimental Setup - Hardware

- Testing two different hardware scenarios for evaluation
- Pre-processing always on worker node

On-board

- All computations on one computing node
- Cluster scenarios:
  - CPU-only setup
  - Each worker has own GPU

Server



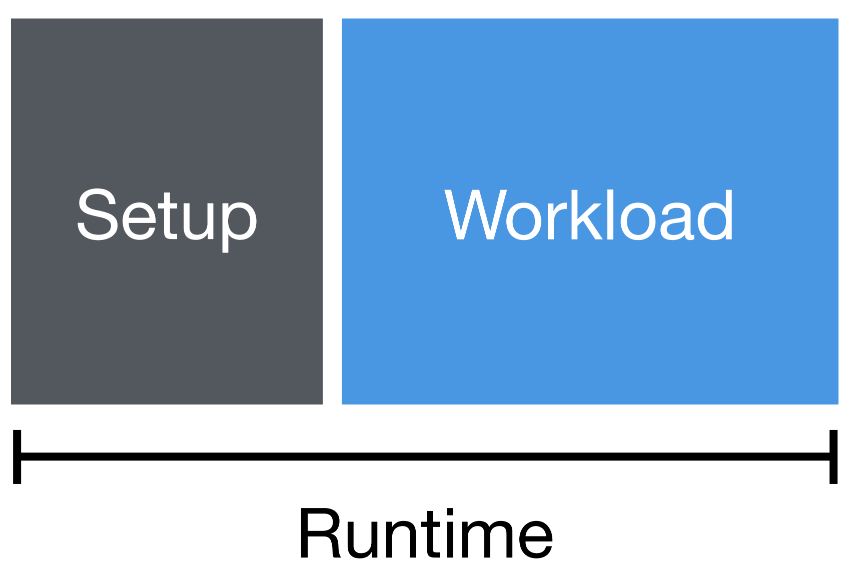
- Graph evaluations on central GPU
- Accessed over network (1Gbit/s)
- Using TensorFlow model server
- Cluster scenario:
  - Not each worker has own GPU

Experimental Setup - Pre-heating

RWTH AACHEN UNIVERSITY
Dennis Noll - 14.02.22

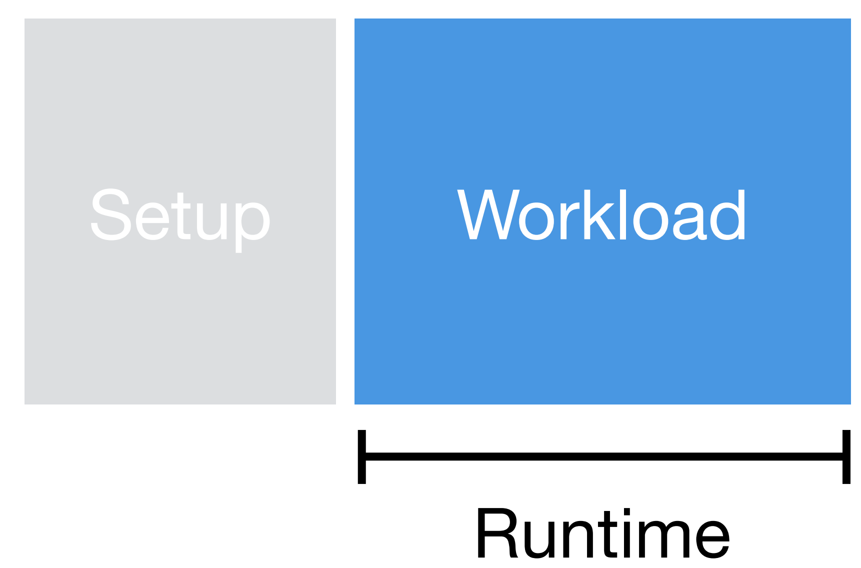
- Graph need to be built before evaluation
- Takes constant time $\mathcal{O}(10s)$
- Testing two scenarios:

Cold



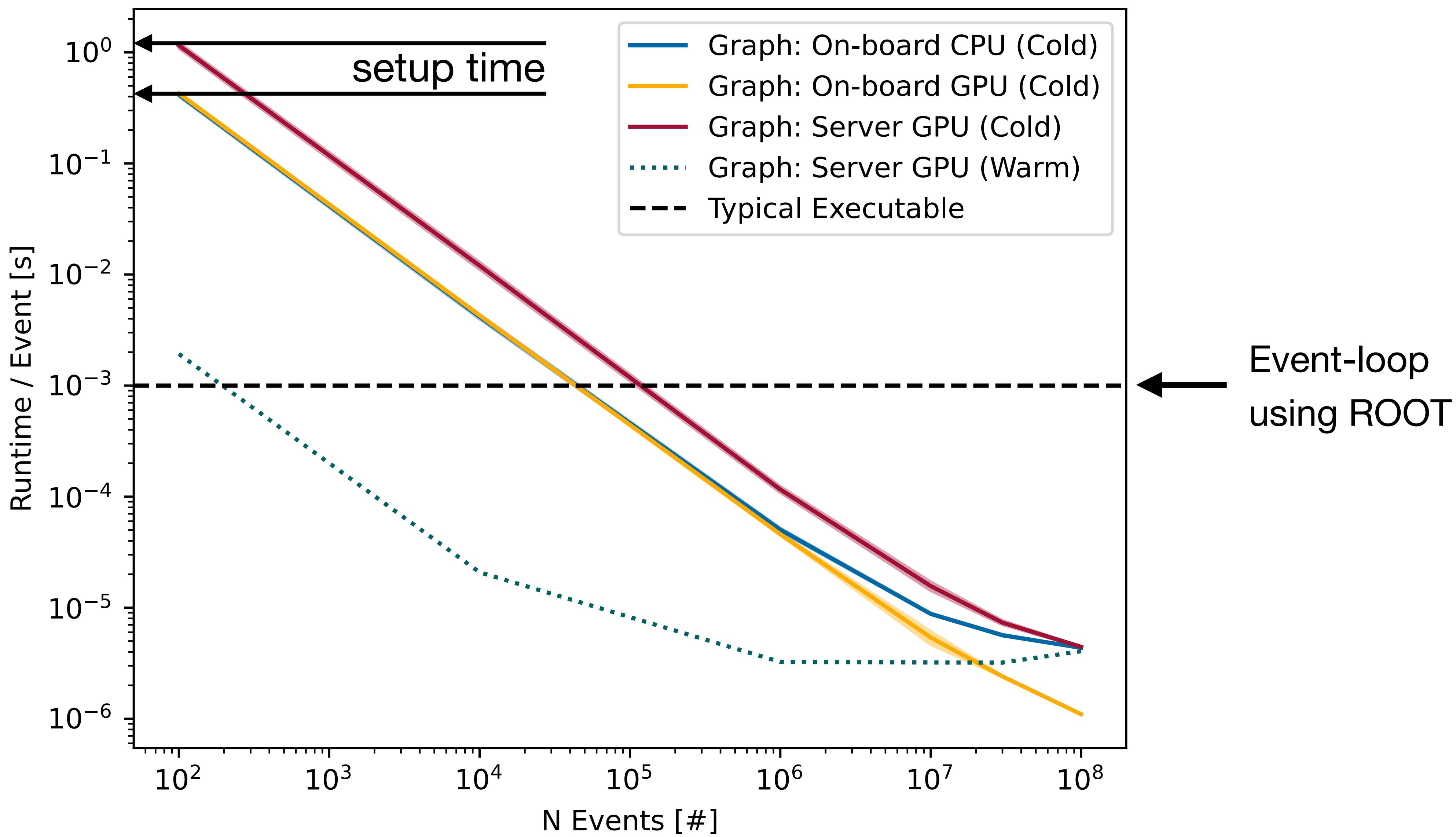
- Setup included in runtime
- Represents:
  - Ad-hoc computation on worker

Warm



- Setup not included in runtime
- Represents:
  - Cluster with central GPU

RWTH AACHEN UNIVERSITY
Dennis Noll - 14.02.22



- For typical analysis (10^8 events):
  - Graph 100x faster than typical executable
  - On-board GPU: Fastest but also most expensive
  - Server GPU: Saturates due to limited network speed

Conclusion

RWTH AACHEN UNIVERSITY
Dennis Noll - 14.02.22

- Future HEP analyses must be fast and resource efficient
- Use vectorised computations and parallelism
- Complex computations parallisable using graph computing models
- Example of neutrino reconstruction:
  - Up to 100x speedup possible for typical analysis

Try on binder
https://bndr.it/8yw3z