Concrete goals and plans for progress

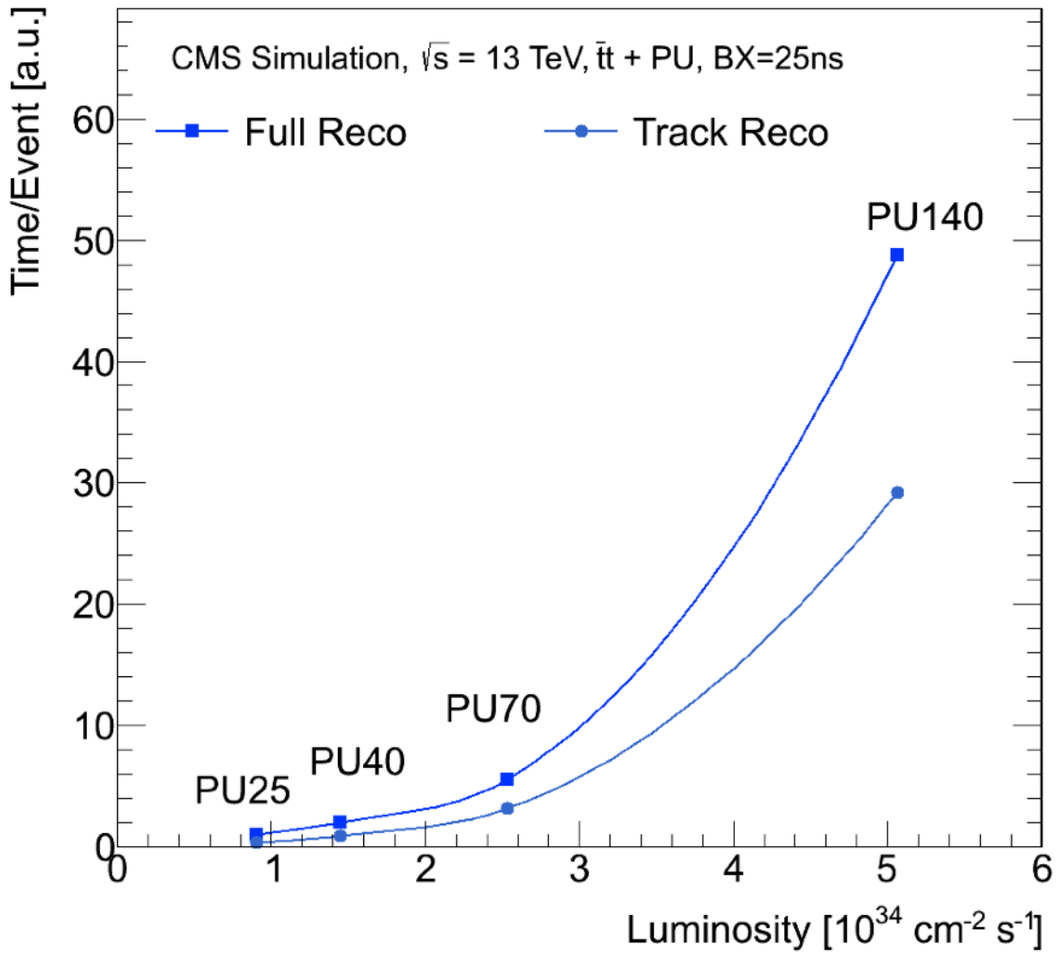
Alexander Schmidt (RWTH Aachen)

2nd Workshop of big data science for (astro)particle physics
21. Feb. 2018

why progress?

we need progress to survive!

- high-lumi LHC will require 60x more CPU resources than 2016 LHC

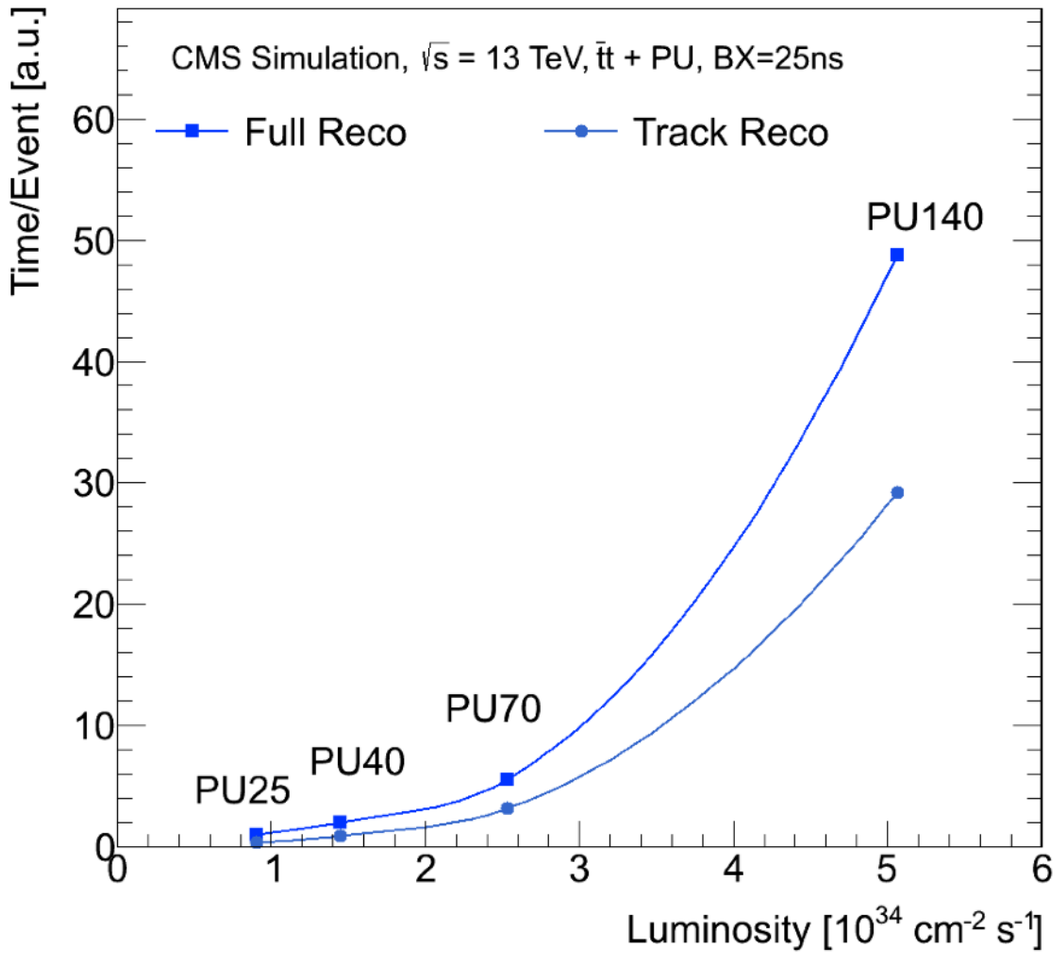


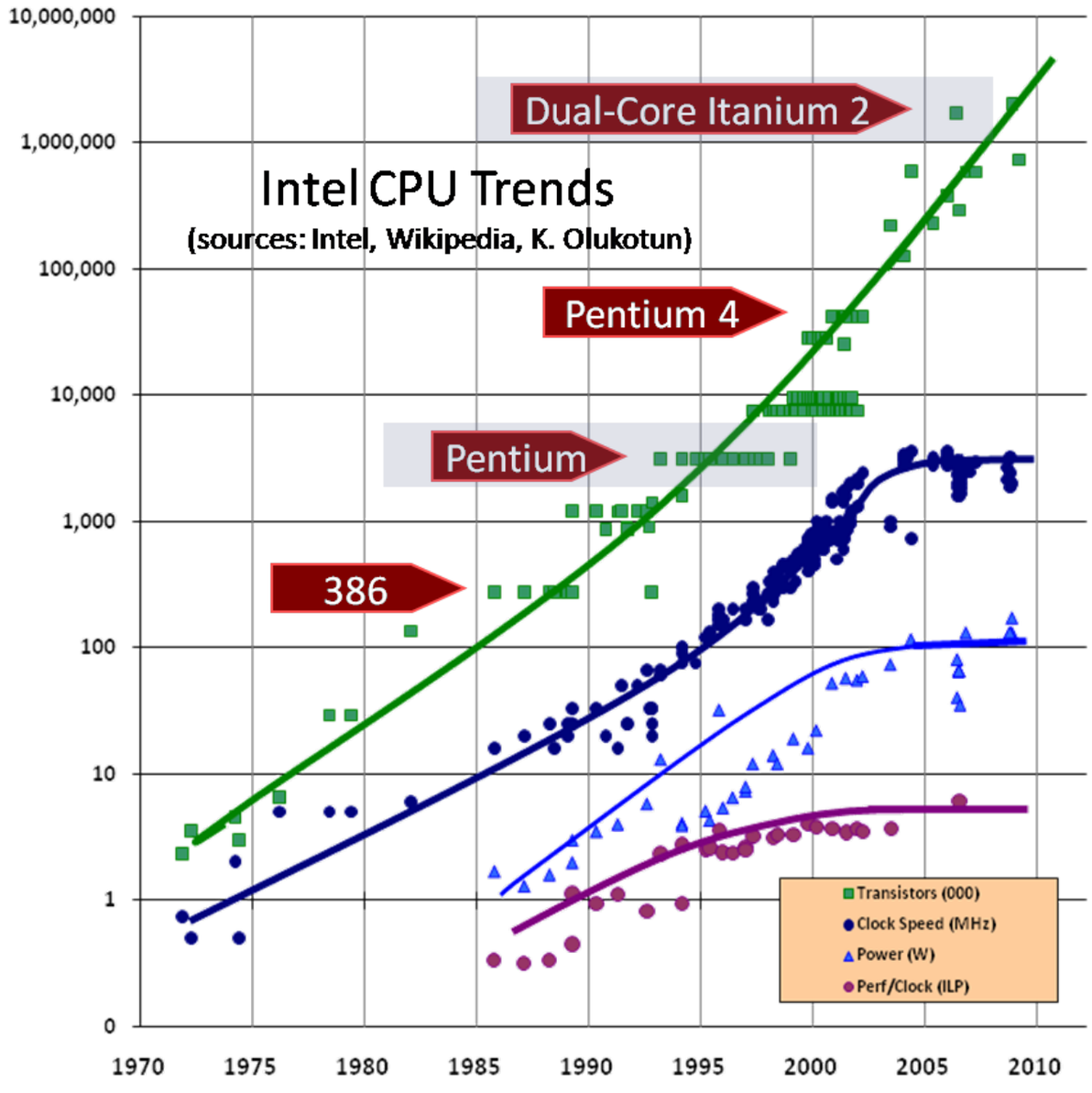
X

why progress?

we need progress to survive!

- high-lumi LHC will require 60x more CPU resources than 2016 LHC





- CPU performance growth has hit the wall
- factor 10 in resources missing to run HL-LHC

X

more to consider

new technologies are emerging and have great success:

- parallel computing
- deep machine learning
- virtualisation
- …

are we (astro/particle physicists) “leading” the developments just as 20 years ago?

- it is now a multi billion EUR industry (google, facebook, amazon)
- we should not be simple customers!
- it’s our job too, to define the field

what is the optimal strategy for our community to benefit?

- workshops like today!
- more workshops?
- a major cross-community strategy?

more to consider

our resources are largely dedicated to physics research

at the same time we need to hugely invest into software/computing

facing all this, we need cross-community effort to

- bundle resources
- do lobbying at funding agencies
- create a common strategy and long-term planning
- develop educational programs
- develop experiment-independent solutions and projects

the plan

all these considerations condensed into a cross-community project plan:

Innovative Digitale Technologien für die Erforschung von Universum und Materie

Gemeinsamer Antrag von Gruppen aus den Bereichen Elementarteilchenphysik, Hadronen- und Kernphysik und Astroteilchenphysik

- Rheinisch-Westfälische Technische Hochschule Aachen, Prof. Dr. Martin Erdmann
- Rheinische Friedrich-Wilhelms-Universität Bonn, PD Dr. Philip Bechtle
- Friedrich-Alexander-Universität Erlangen-Nürnberg, Prof. Dr. Gisela Anton
- Goethe Universität Frankfurt am Main, Prof. Dr. Volker Lindenstruth
- Albert-Ludwigs-Universität Freiburg, Prof. Dr. Markus Schumacher
- Georg-August-Universität Göttingen, Prof. Dr. Arnulf Quadt
- Universität Hamburg, Jun.-Prof. Dr. Gregor Kasieczka
- Karlsruher Institut für Technologie, Prof. Dr. Günter Quast
- Johannes Gutenberg-Universität Mainz, Prof. Dr. Volker Büscher
- Ludwig-Maximilians-Universität München, Prof. Dr. Thomas Kuhr
- Bergische Universität Wuppertal, Prof. Dr. Christian Zeitnitz

Assoziierte Partner sind

- CERN, Dr. Markus Elsing
- DESY, Dr. Volker Gülzow
- GridKa, Dr. Andreas Heiss
- GSI Helmholtzzentrum für Schwerionenforschung, Darmstadt, Dr. Kilian Schwarz
- Forschungszentrum Jülich, Dr. Elisabetta Prencipe
- Westfälische Wilhelms-Universität Münster, PD Dr. Christian Klein-Bösing

submitted to BMBF
October 2017

X

four main work areas

A: developments for the use of heterogeneous computing resources

B: application and test of virtualised software components

C: deep learning

D: event reconstruction: cost- and energy efficient use of computing resources

work area A: heterogeneous computing resources

- **goal: mitigate resource gap through access to heterogeneous resources**
  - opportunistic resources (free capacities on HPC centers)
  - embed “volunteer computing”, scientific and commercial cloud services
  - adapt existing grid, batch systems transparent to the user

work area A: heterogeneous computing resources

- **goal: mitigate resource gap through access to heterogeneous resources**
  - opportunistic resources (free capacities on HPC centers)
  - embed “volunteer computing”, scientific and commercial cloud services
  - adapt existing grid, batch systems transparent to the user

A1) Werkzeuge zur Einbindung • Scheduling von Cloud-Jobs • Container-Technologien • Checkpointing • Zugang zu Experiment-Datenbanken	**A2) Effiziente Nutzung** • Transiente Datencaches • Transparenter Zugriff auf verteilte Daten
A3) Workflow-Steuerung • Identifikation und Steuerung • In-Pilot Job-Monitoring • Accounting • Optimierung durch Data-Mining	

X

work area B: virtualisation

much related to A:

- large scale deployment of virtualisation, e.g. through containers
- “infrastructure as a service”
- interoperability of components

X

work area B: virtualisation

much related to A:

- large scale deployment of virtualisation, e.g. through containers
- “infrastructure as a service”
- interoperability of components

B1) Tests der Technologiekomponenten Implementierung und Tests auf verschiedenen Plattformen von • Speicher- und Cachinglösungen und • virtualisierter Dienste (Datenbanken, Monitoring, Accounting).	**B2) Job- und Ressourcenmanagement** Jobverteilung und Überwachung in der Umgebung heterogener Computingressourcen unter Einbeziehung von Containervirtualisierung.
B3) Virtualisierung von Nutzerjobs • Erfassung der Anforderungen, • Bestimmung und Erzeugung der Laufzeitumgebung, • Erstellung des Containers und von Metadaten und • Checkpointing von Containervirtualisierung.	**B4) Kombinierte Tests** Testen von Gesamtsystemen (Speicher, Dienste, Ressourcemanagement) auf verschiedenen Plattformen in Bezug auf • Installations- und Wartungsaufwand, • Performance, • Skalierbarkeit und • Robustheit.

X

work area C: deep learning

- application of deep learning in (astro)particle physics
- establishment of community wide structures (synergy)

X

work area C: deep learning

- application of deep learning in (astro)particle physics
- establishment of community wide structures (synergy)

C1) Sensornahe Verarbeitung von Daten • Signalfilter, Rauschunterdrückung • Verarbeitung von zeitabhängigen Signalen	**C2) Objektrekonstruktion** • Spur- und Clusterrekonstruktion, Jetbildung, Ereignisrekonstruktion • Fragestellungen für Anordnung, Reihenfolge, Zuordnungen von Daten • Optimierungen zur Extraktion kleiner Signale bei großem Untergrund
C3) Netzwerkbeschleunigte Simulationen • Generative adversarial networks, Anpassung von Simulationen an Datenverteilungen • Evaluationsverfahren für die Qualität der Netzwerksimulationen	**C4) Qualität von Netzwerkvorhersagen** • Reduzierung experimenteller systematischer Unsicherheiten • Spezielle Lernstrategien • Vorhersagenrelevante Information • Unsicherheiten von Vorhersagen

C1: sensor-level data

- **explore usability of deep learning methods at sensor level:**
  - filtering of signals, noise suppression
  - large applicability: calorimeter cells, tracker hits, radio signals, photo sensors

  - time-dependent signals:
    - often not fully exploited
    - extract information hidden in sequence of signal development

C1: sensor-level data

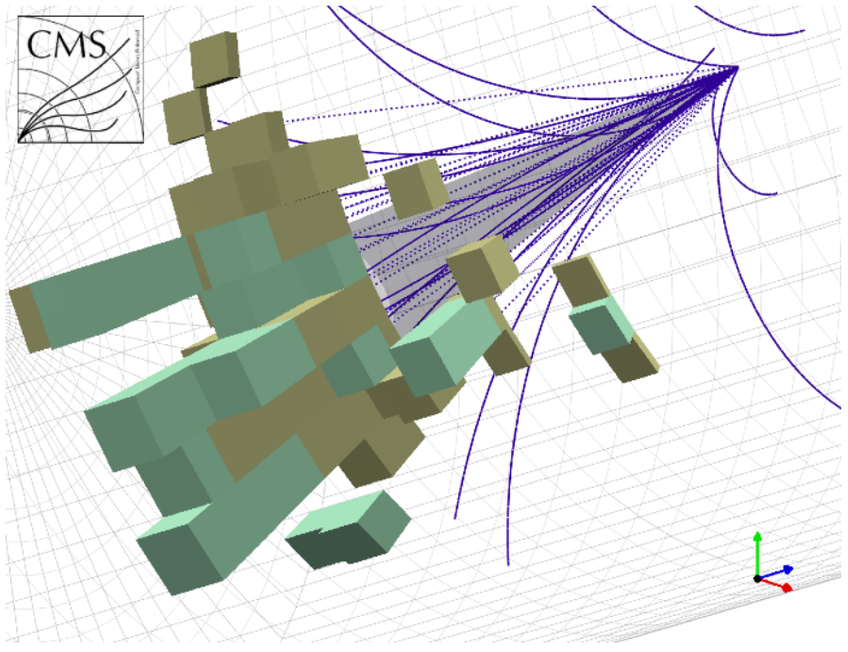
- **explore usability of deep learning methods at sensor level:**
  - filtering of signals, noise suppression
  - large applicability: calorimeter cells, tracker hits, radio signals, photo sensors

  - time-dependent signals:
    - often not fully exploited
    - extract information hidden in sequence of signal development

- **concrete project (Aachen):**
  - air-shower data at Pierre Auger radio array
  - difficult signal/noise
  - analyse signal time-development, amplitude, shape
  - test applicability of architectures like autoencoder

C2: object reconstruction

- **use deep learning in a wide range of physics objects:**
  - tracks from tracking chamber hits
  - clusters from calorimeter signals
  - jets from clusters
  - properties of jets (flavour, charge, origin,…)
  - very successful already in the past, promising for future
  - concrete projects:
    - boosted jets (Hamburg),
    - classification of active galactic cores (Blazars) through EM spectrum (Erlangen)



X

C3: simulation

- **use deep networks for simulation:**
  - simulation of calorimeter showers and air showers is extremely expensive
  - adversarial networks can be used for simulation
  - need to understand where it is useful
  - develop evaluation criteria for quality of simulation

  - concrete projects:
    - implement interaction of radiation with matter in a generative network to obtain full instantaneous shower simulation (Karlsruhe)
    - mass production of simulated datasets via generative networks, development of filter criteria (Munich)

C4: quality of prediction

- **goal: define catalog of criteria for usability of DNN in physics analyses**
  - which information is actually used by DNN? which information is needed?
  - reduce impact of systematic effects (e.g. adversarial training)
  - stability/robustness of predictions, causality
  - exploit already existing tools (PatternNet, PatternLRP)

work area D

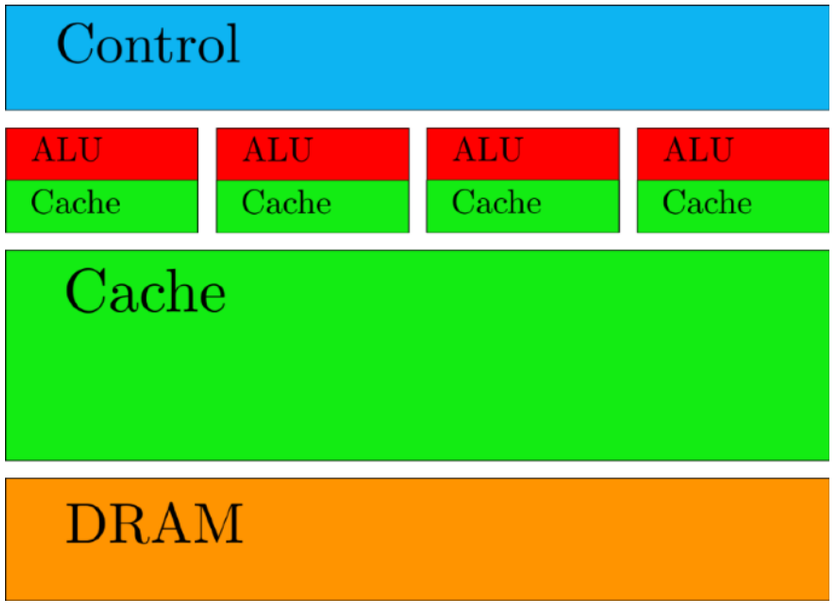
goals:

- adapt (astro)particle software to modern computing architectures (GPUs)
- make software more resource-efficient
- develop experiment-independent libraries

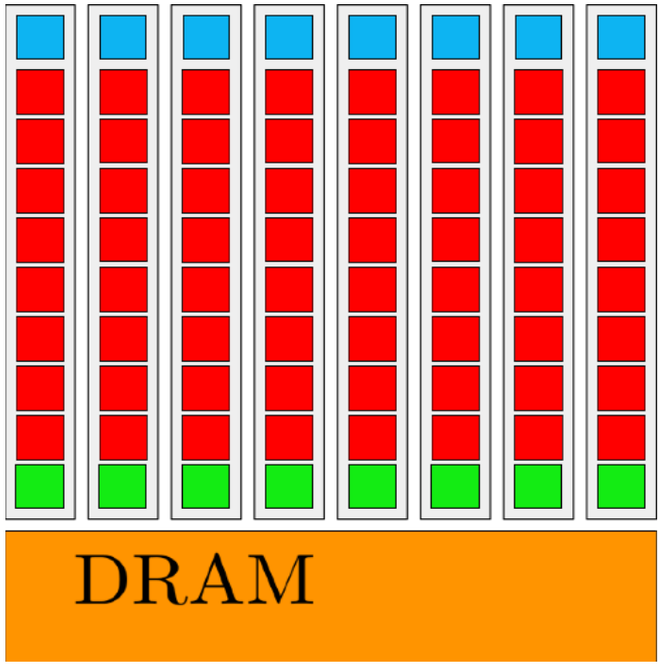
D1) Spurfindung • alternative Algorithmen, z.B. zellulärer Automat • alternative Architekturen, z.B. GPUs	**D2) Parameterbestimmung** • Verknüpfung GenFit2-ACTS
D3) Neutrinoexperimente • dünnbesetzte Detektorinstrumentierung • variable Signalzeit als kritische Information	

X

example: parallelism



CPU



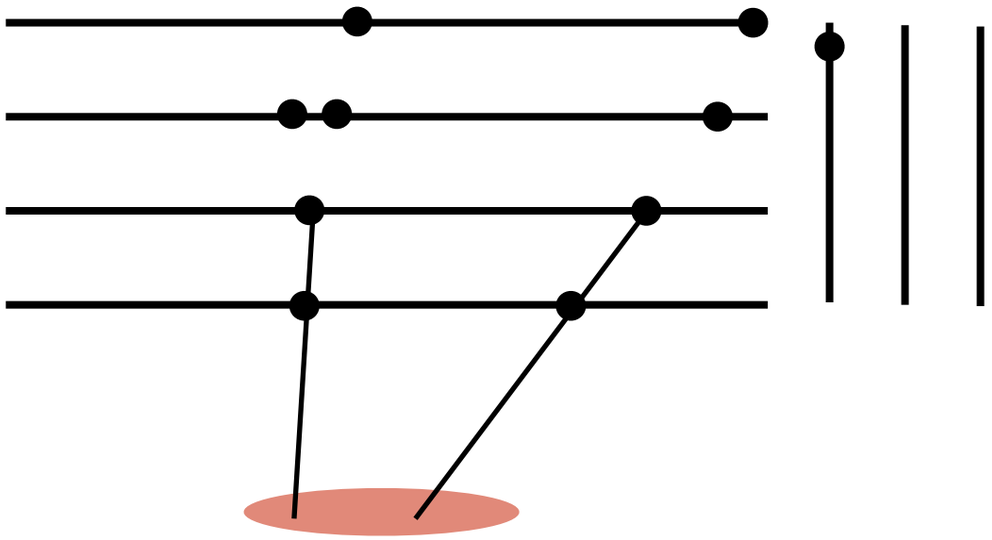
GPU

X

example: parallelism

fourth detector layer

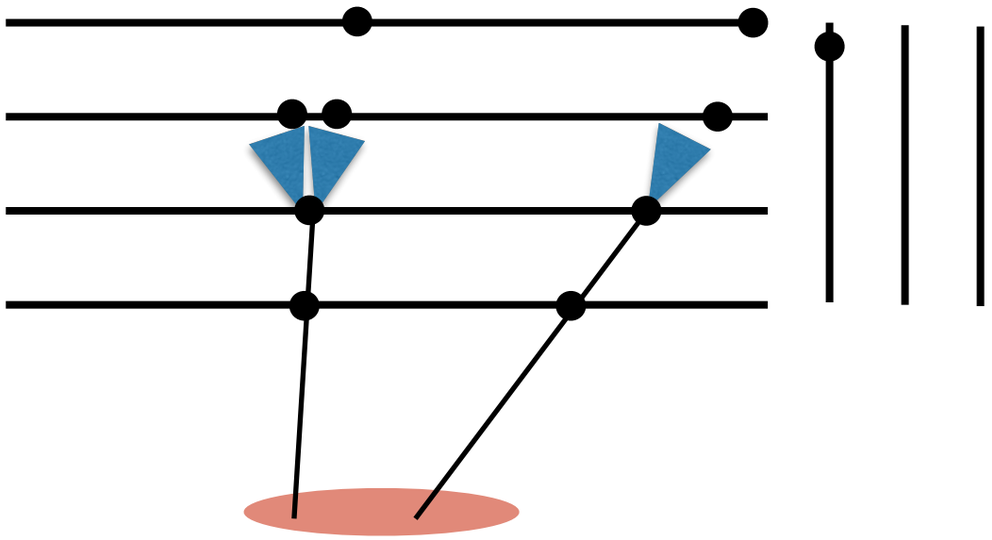
- quadruplet seeding: propagate triplets to fourth layer
- natural extension of current algorithm



example: parallelism

fourth detector layer

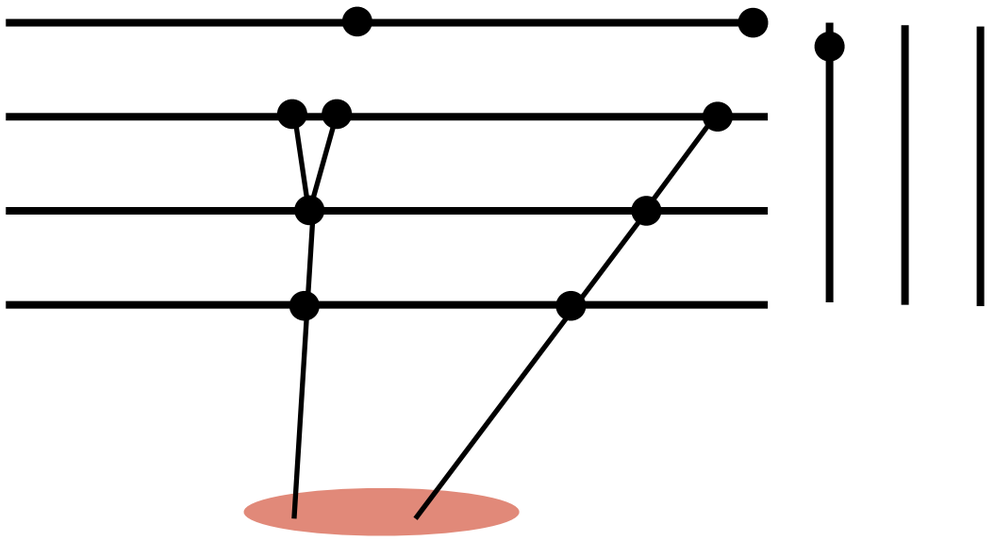
- quadruplet seeding: propagate triplets to fourth layer
- natural extension of current algorithm



example: parallelism

fourth detector layer

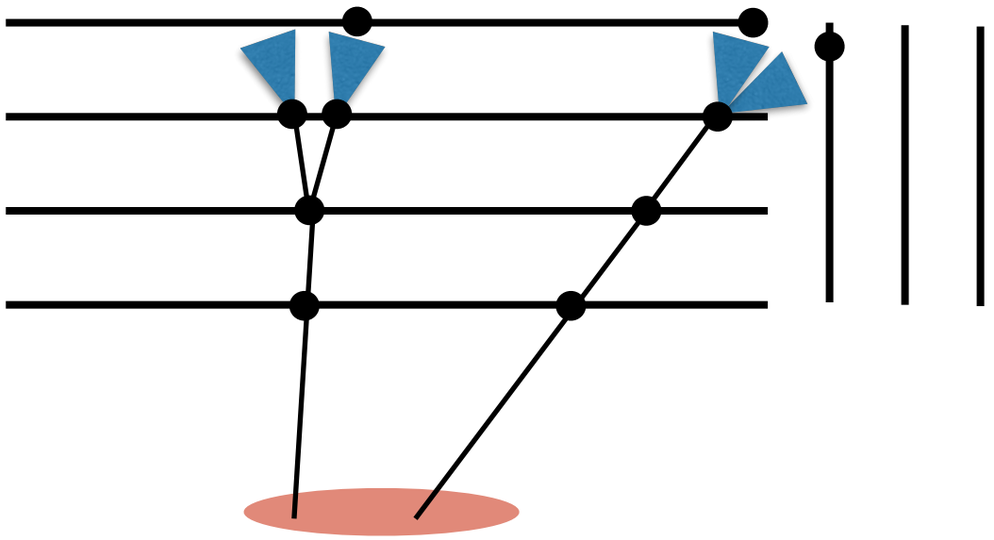
- quadruplet seeding: propagate triplets to fourth layer
- natural extension of current algorithm



example: parallelism

fourth detector layer

- quadruplet seeding: propagate triplets to fourth layer
- natural extension of current algorithm

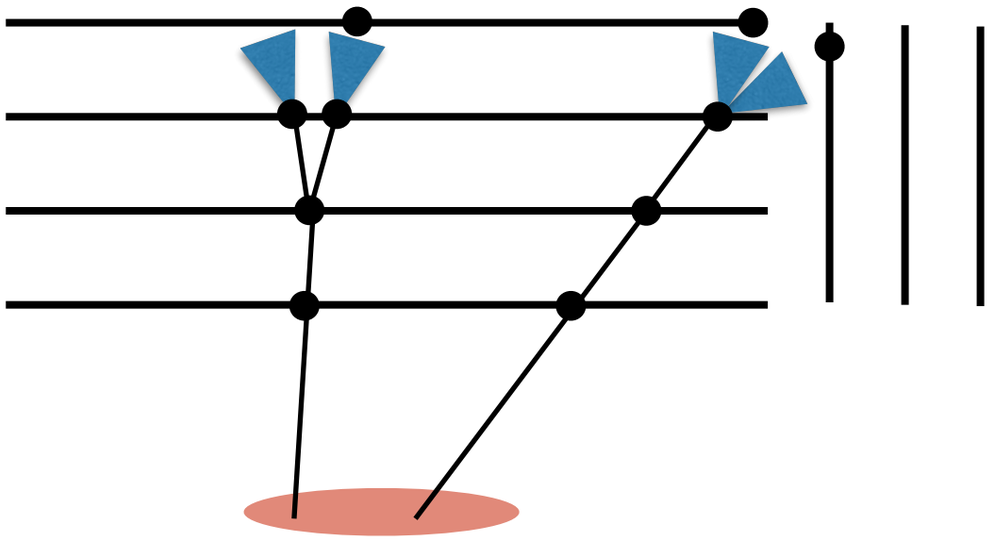


- computing time grows exponential with PU

fourth detector layer

- quadruplet seeding: propagate triplets to fourth layer
- natural extension of current algorithm

- new seeding algorithm
- based on parallel-friendly algorithmic structure (cellular automaton)

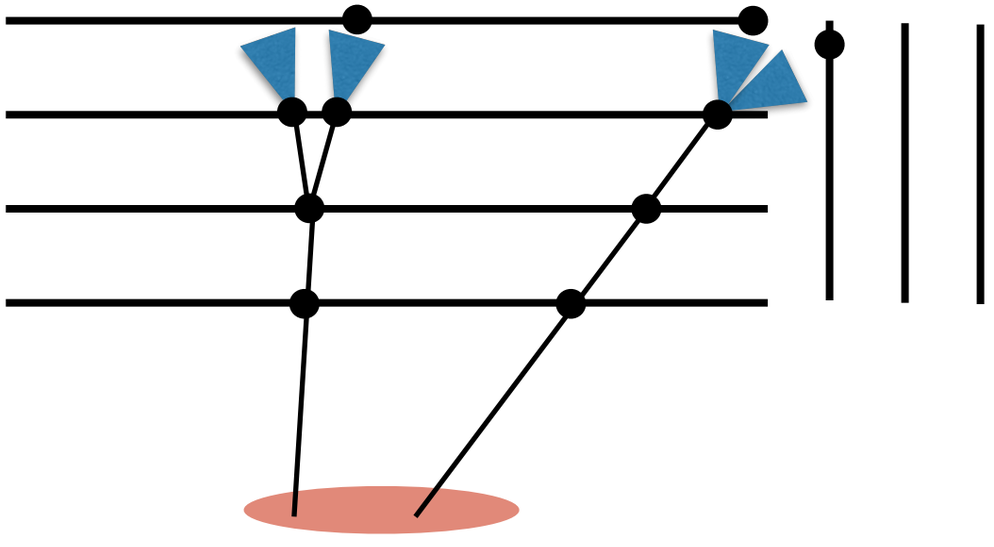


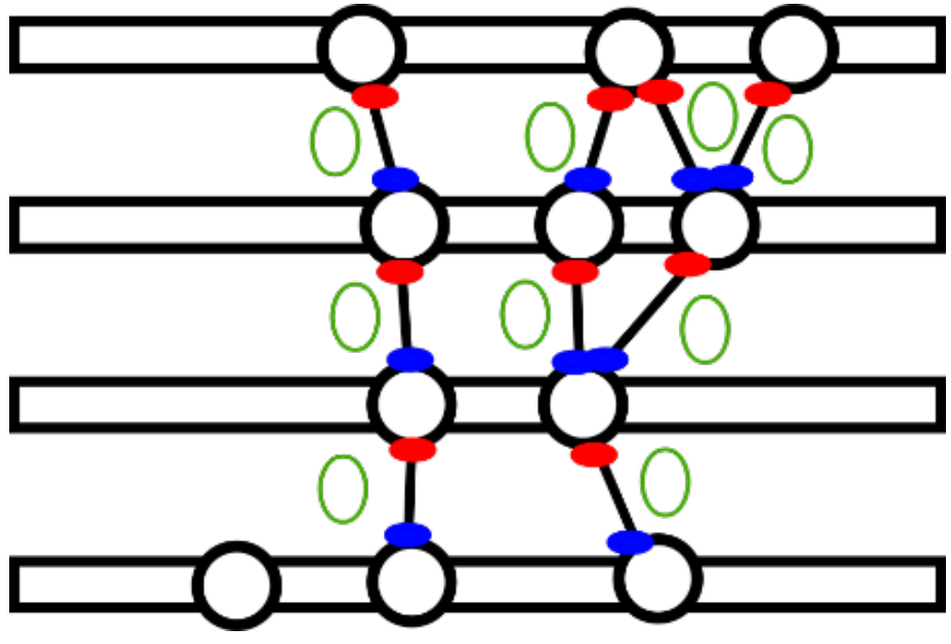
- computing time grows exponential with PU

fourth detector layer

- quadruplet seeding: propagate triplets to fourth layer
- natural extension of current algorithm

- new seeding algorithm
- based on parallel-friendly algorithmic structure (cellular automaton)

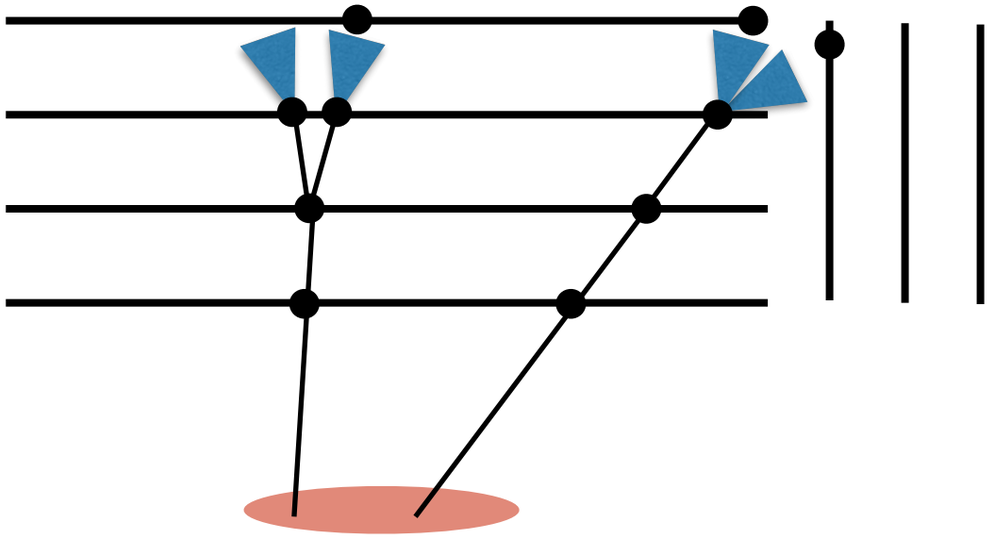




- computing time grows exponential with PU

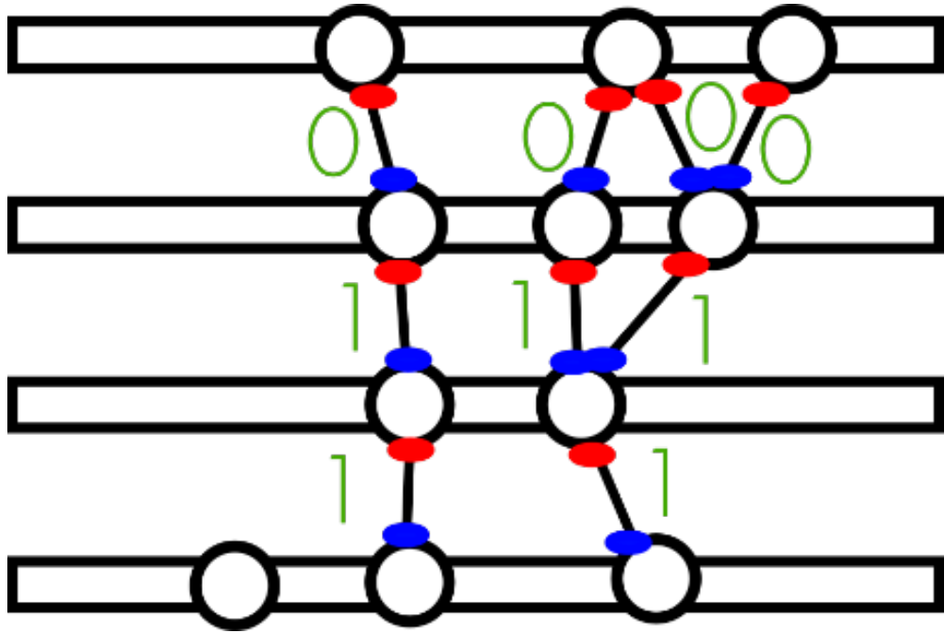
fourth detector layer

- quadruplet seeding: propagate triplets to fourth layer
- natural extension of current algorithm



- computing time grows exponential with PU

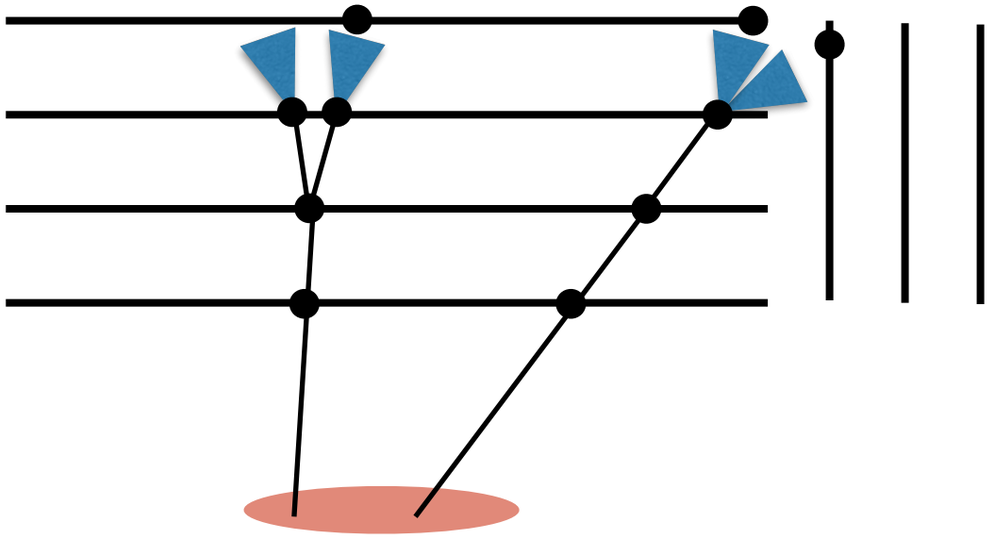
- new seeding algorithm
- based on parallel-friendly algorithmic structure (cellular automaton)



example: parallelism

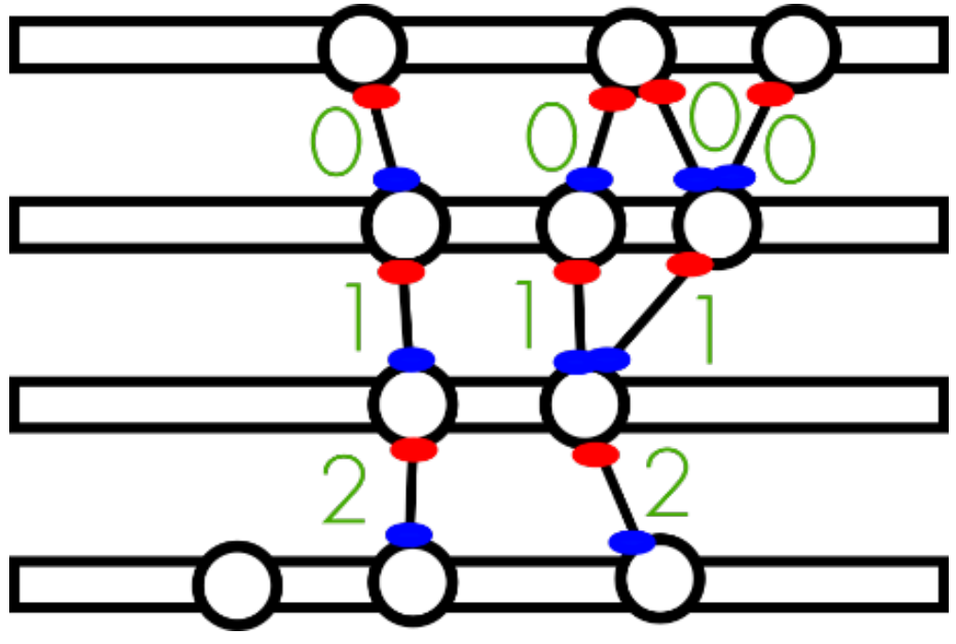
fourth detector layer

- quadruplet seeding: propagate triplets to fourth layer
- natural extension of current algorithm



- computing time grows exponential with PU

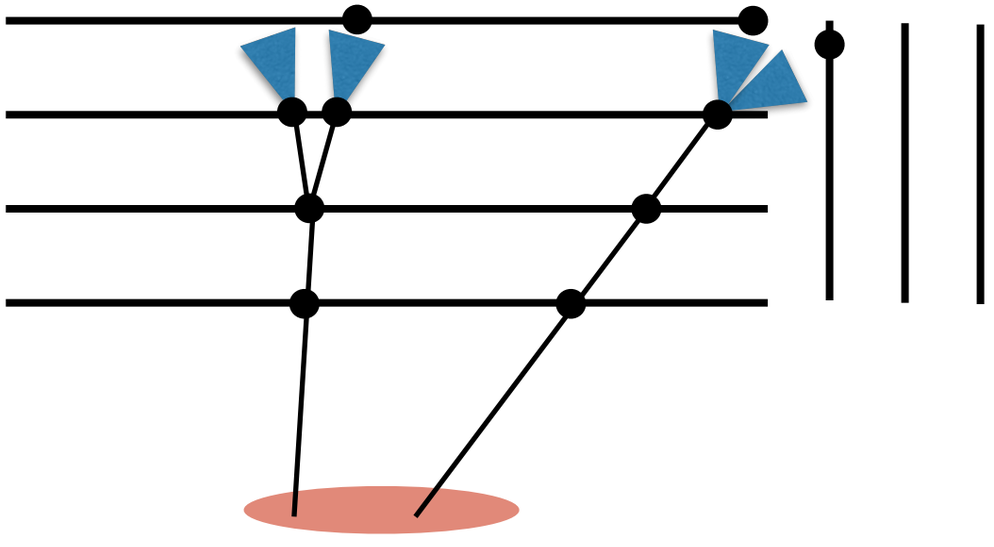
- new seeding algorithm
- based on parallel-friendly algorithmic structure (cellular automaton)



example: parallelism

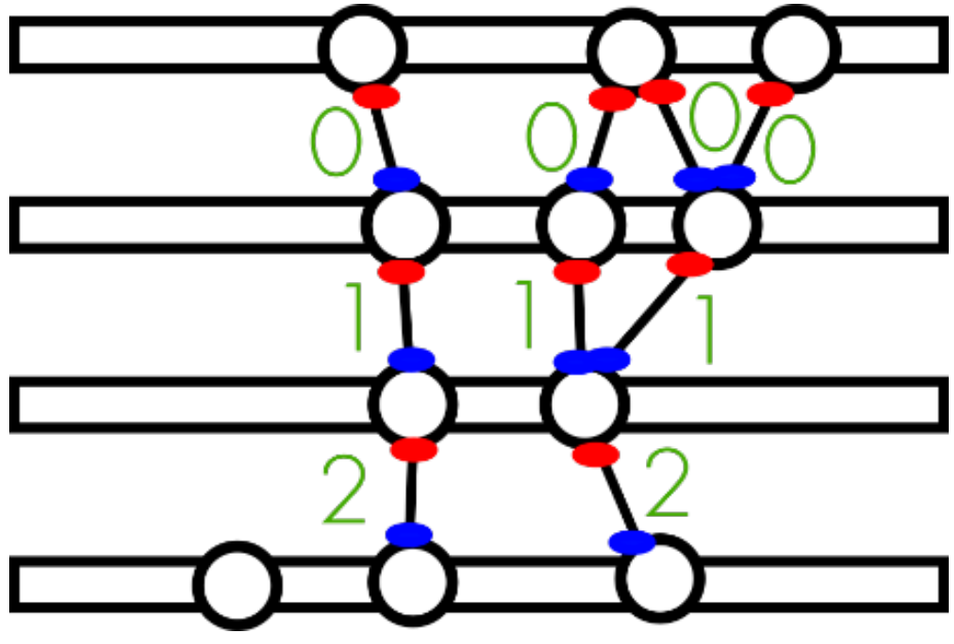
fourth detector layer

- quadruplet seeding: propagate triplets to fourth layer
- natural extension of current algorithm



- computing time grows exponential with PU

- new seeding algorithm
- based on parallel-friendly algorithmic structure (cellular automaton)



- computing time grows linear with PU

example: parallelism

	time per event CPU (ms)	time per event GPU (ms)
Triplet propagation	66.3	
cellular automaton		1.6

- Hardware used:
  - CPU Intel 4771K
  - GPU NVIDIA K40

example: parallelism

	time per event CPU (ms)	time per event GPU (ms)
Triplet propagation	66.3	
cellular automaton	22	1.6

- Hardware used:
  - CPU Intel 4771K
  - GPU NVIDIA K40

example: parallelism

	time per event CPU (ms)	time per event GPU (ms)
Triplet propagation	66.3	
cellular automaton	22	1.6

- Hardware used:
  - CPU Intel 4771K
  - GPU NVIDIA K40

- **improve** physics performance at the same cost (easily run tracking on all events in HLT)

- or **save** millions of EUR at same physics performance

example: parallelism

	time per event CPU (ms)	time per event GPU (ms)
Triplet propagation	66.3	
cellular automaton	22	1.6

- Hardware used:
  - CPU Intel 4771K
  - GPU NVIDIA K40

- **improve** physics performance at the same cost (easily run tracking on all events in HLT)

- or **save** millions of EUR at same physics performance

→change our approach to algorithm development

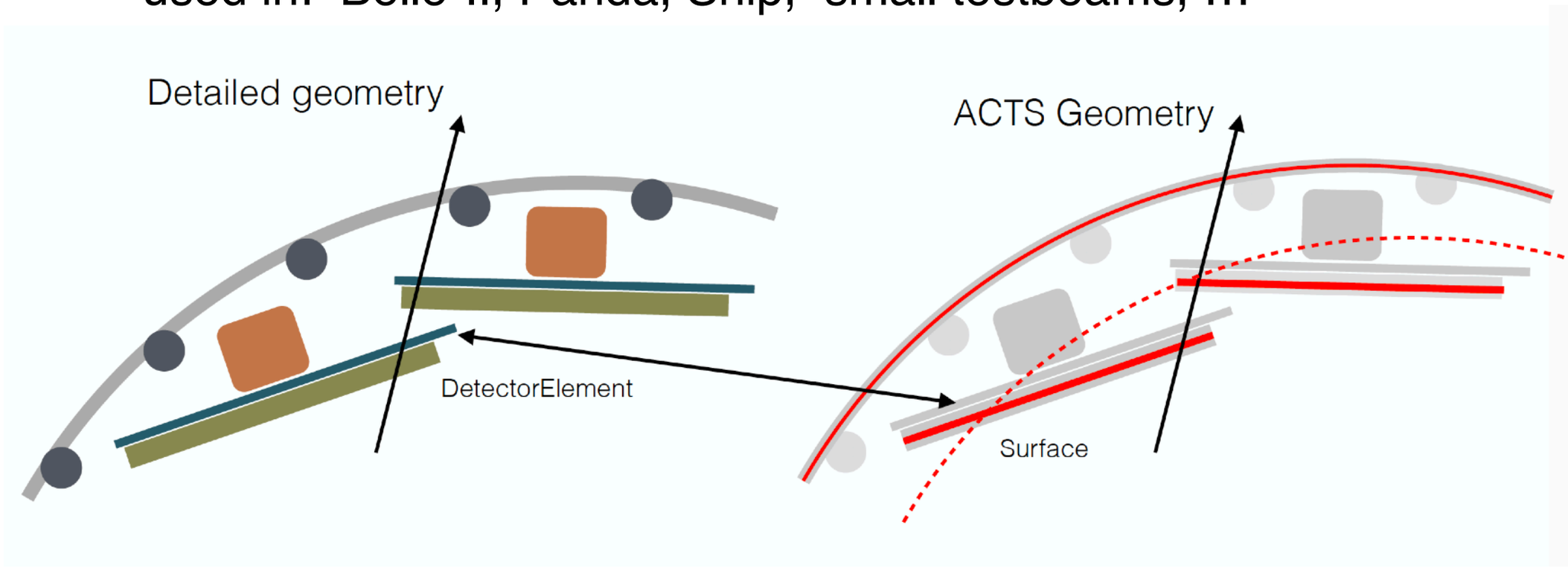
example: parallelism

- **successfully running in CMS since 2017**
- **developments continue (e.g. full tracker)**
- **being ported to other experiments (e.g. LHCb)**
- **future: experiment-independent libraries**

example: GenFit2-ACTS

ACTS:

- started from ATLAS track reconstruction software
- aims at encapsulating the reconstruction into generic experiment-independent package
- provides high-level data structures and algorithms
- alternative: GenFit2
- used in: Belle-II, Panda, Ship, small testbeams, …



- concrete projects:
  - investigate usability of ACTS in non-HEP experiments. Use ACTS in GenFit2 backend. Understand the limitations and adapt the software accordingly (Karlsruhe)
  - adapt ACTS for ILC (DESY)
  - adapt ACTS for Panda (FZJ)

summary

- **computing/software in (astro)particle physics is the key to survival and the driver of fantastic new opportunities**

- **“local efforts” not sufficient, community-wide coordination has been initiated**

- **writing this community-wide project plan already established excellent communication links**

- **this is an iterative process and needs YOUR input**