Software for PED studies

(About) Compute Resource Needs

7th FCC Physics Workshop 2024
Annecy, France

January 31, 2024
G Ganis, CERN-EP

Outline

- Quick recap about the problem
- How it looks for FCC-ee
- A few remarks

Computing and HEP

- Cost of IT-related components in HEP experiments, from software development to storage to processing power, constantly raised in the last decades
  - Starting with 90’s, i.e. when computing has become significant
- Experiment workflows and their components have become more and more complex in many directions
  - Increasing expected data samples, hence increasing needs
  - Varying scenario of ressources
    - From single-core to multi-core to heterogeneous processing units
    - Evolution of storage systems from local to distributed and cache hierarchies
  - Continuous evolution/optimisation of software and data structures
- Estimating/predicting needs for computing resources has become crucial to plan for and secure them
- All this requires modeling

Modeling the resource needs

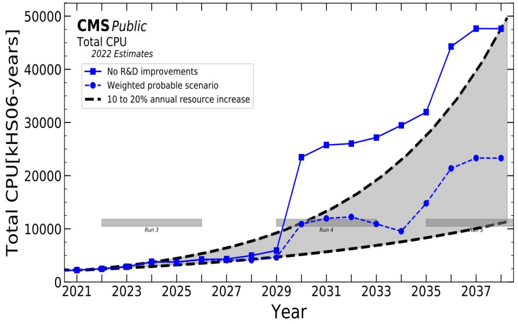
- Simple, in principle
  - Define the needs and the activities to satisfy them
    - E.g. ‘MC samples for year Y_N’ requires the activities
      - Event generation, simulation, reconstruction
  - To each given activity corresponds a set of workflows, each defining a set of resources to be used, e.g.
      - Event generation on GPU producing output on local NAS
      - Simulation on the Grid producing output on the Grid
      - Reconstruction on HLT producing output on EOS
  - Put everything together to get the compute ressources needs
    - Taking into account experiment guidelines and policies

- All this depends on assumptions which may be or may be not well defined
- For LHC they reasonably well defined

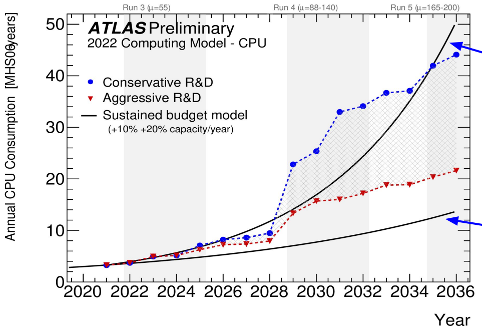
Inspired by D Lange et al, [CMS Computing Resources Modeling]

Projections of resource needs of HL-LHC

Processing power



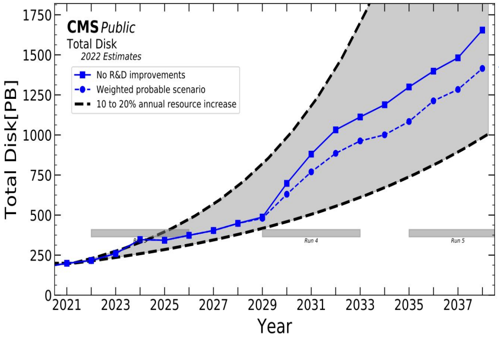
Better software



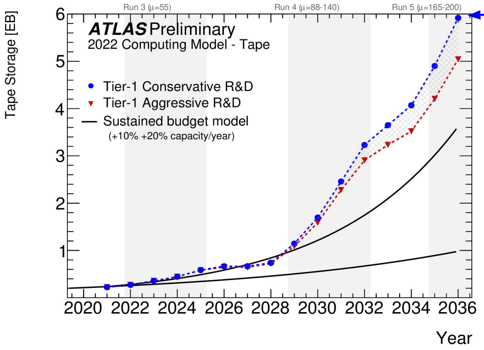
Pledged +20%

Pledged +10%

Storage



Better software



6 EB

CMS, ATLAS

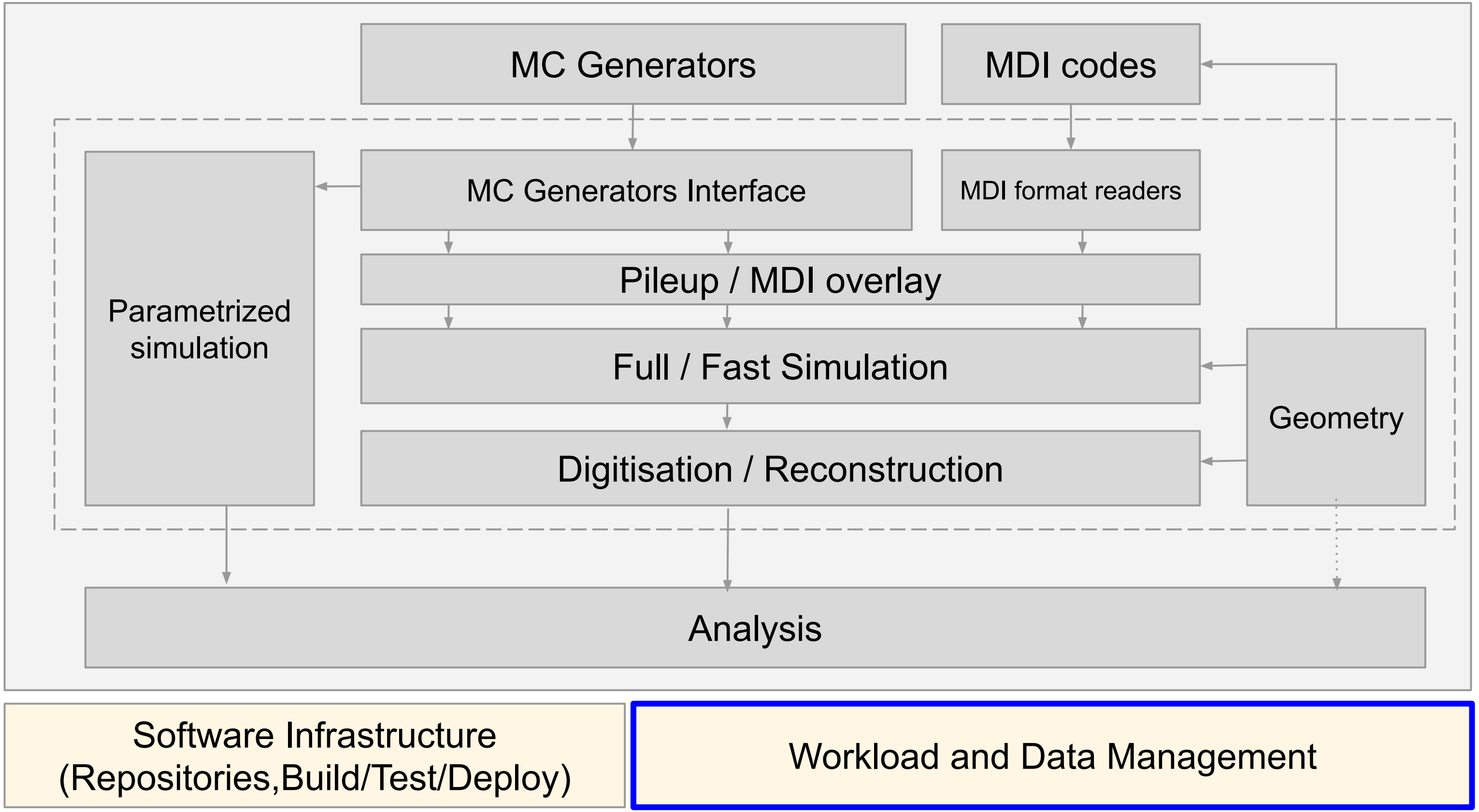
How does it look for FCC-ee?

Modeling resources for a project “en devenir”

- Same reasoning, less precise assumptions, different purpose
  - Monte Carlo only
  - No data processing, calibration, …
  - Several detector concepts
  - Several digitisation, reconstruction options, …

- Several purposes: e.g.
  - Projects needs for pledged resources “in production”
  - Estimate the potential of a limited set of resources to set priorities in the short term

Workflows to support for FCC



What we have now

- CERN
  - EOS volumes
    - 500 TB for central productions (157 TB free, stil used by some CDR files)
    - 200 TB for analysis, starts to be used
  - CPU: 9000 HS06 on lxbatch
  - Integrated in iLCDirac
- Other sites integrated
  - BARI
  - CNAF
  - Glasgow (storage only)
- Some GPU resources
  - CERN, EuroHPC

Not yet limited, in general, but start reaching the boundaries

A first resources analysis for FCC-ee[1]

- Assumptions
  - Nominal luminosities
    - {90, 12, 5, 0.2, 1.5} ab^{-1} at $\sqrt{s}$ = {91.2, 160, 240, 350, 365} GeV
  - MC reference sample = data sample
    - i.e. 3×10^{12} visible Z decays, 10^8 WW events, 10^6 ZH events, 10^6 tt events
- Event sizes (see next)
  - RAW: 1 - 2 MB/evt
  - AOB: 5 - 10 kB/evt
- Processing Power
  - CERN Openstack Core = 10-15 HEPSpec06
  - FCC currently assigned processing units = Computing Unit = 9000 HEPSpec06
  - CERN OpenStack node used for tests: 16 cores, 32 GB RAM

1. Based on: GG, C Helsens: EPJ Plus (2022) 137:30

Event Sizes estimations[1]

Table 2 Typical RAW event sizes in kB for the **Z** run for the two baseline detector solutions [12] and the ALEPH detector [13]; the contribution of the final states originating from the Z exchange (Z decays) is singled out from the expected total (all events)

Readout channels	CLD		IDEA		ALEPH	
	1.9 G		2.8 G		1 M	
Sub-detector	Z decays (kB)	All events (kB)	Z decays (kB)	All events (kB)	Z decays (kB)	All events (kB)
Vertex	1.3	62	1.3	62		
Tracker	1.4	102	500	595		
Calorimeter	230	920*	500	2000*		
Muon	0.03	0.75	0.03	0.75		
Total	233	1085	1001	2658	120	550

RAW: 1 - 2 MB / evt

AOD: 5 - 10 kB / evt

Table 3 Typical FCC-ee event sizes, in kB, for different types of events processed through DELPHES in EDM4hep format (using the IDEA detector concept card with track covariance [14])

Process	$\sqrt{s}$ (GeV)	Average size (kB)
$Z \to u\bar{u}, d\bar{d}, s\bar{s}$	91.2	4.9
$Z \to c\bar{c}$	91.2	5.2
$Z \to b\bar{b}$	91.2	5.5
$Z \to \tau^+\tau^-$	91.2	1.2
Z decays, ALEPH, AOD	91.2	9.0
Z decays, ALEPH, MINI	91.2	2.2
ZH inclusive	240	8.9
ZZ inclusive	240	6.6
W^+W^- inclusive	240	6.4
$t\bar{t}$	350	13

1. Based on: F. Grancagnolo: Event Rates at Z-pole, talk presented at 4th FCC PED workshop, Nov 2020

Storage requirements

Table 4 RAW data estimates for FCC-ee

Run	$\sqrt{s}$ (GeV)	Statistics	RAW data
Z	91.2	3×10^{12} Z decays (visible)	3–6 EB
WW	160	10^8 W^+W^- events	0.1–0.2 PB
ZH	240	10^6 ZH events	1–2 TB
$t\bar{t}$	350, 365	10^6 $t\bar{t}$ events	1–2 TB

≍HL-LHC

Table 5 AOD data estimates for FCC-ee

Run	$\sqrt{s}$ (GeV)	Statistics	AOD data
Z	91.2	3×10^{12} Z decays (visible)	15–30 PB
WW	160	10^8 W^+W^- events	0.5–1 TB
ZH	240	10^6 ZH events	5–10 GB
$t\bar{t}$	350, 365	10^6 $t\bar{t}$ events	5–10 GB

Computing requirements

Table 6 Time estimated to generate $q\bar{q}$, $\tau^+\tau^-$ and $\mu^+\mu^-$ events at the Z peak

Generator	Process	100k/core (s)	**Z** sample/core	**Z** sample/9000 HS06 (days)
Pythia8	$q\bar{q}$	148	4×10^9 s = ~126 y	50–75
KKMCee	$q\bar{q}$	151	4×10^9 s = ~126 y	50–75
KKMCee	$\tau^+\tau^-$	195	0.25×10^9 s = ~8 y	3–4.5
KKMCee	$\mu^+\mu^-$	334	0.44×10^9 s = ~14 y	5–7.7

Table 7 Time estimated to simulate $q\bar{q}$ events at the Z peak

Process	1k/core	**Z** sample/core	**Z** sample/9000 HS06
$q\bar{q}$	20k s = 5 h 33 min	6×10^{13} s = ~ 1.9×10^6 y	$2.1–3.2 \times 10^3$ y

2000-3000 years!

Table 8 Time estimated to simulate $q\bar{q}$ events at the Z peak with DELPHES

Process	100k/core	**Z** sample/core	**Z** sample/9000 HS06
$q\bar{q}$	212 s	6.4×10^9 s = ~ 202 y	0.22–0.34 y

Computing estimation remarks

- MC event generator can be challenging for full-scale production
  - Code optimisations and/or filtering techniques might be required
- Full simulation times have been cross-checked with ATLAS times for similar multiplicity
  - Recent Geant4 is up to a factor of 2 faster for ATLAS
  - Fast simulation techniques (see A Zaborowska talk) might help
  - Other option: selected simulation, i.e. non simulating what is never touched is not simulated
- Reconstruction
  - Between 10% (ALEPH) - 30% (BELLE) of simulation
  - Could really benefit from using heterogenous resources

Putting all together ...

- The FCC-ee resource needs are of the same order of HL-LHC
  - If HL-LHC solves the problems, FCC-ee gets it for free

- Putting all together

Table 9 Number of $q\bar{q}$ events that can be produced per day with one computing unit and with the equivalent of the ATLAS computing resources

	Generation	Simulation	Reconstruction	DELPHES
Computing unit	$3.5–5.2\times10^{10}$	$2.6–3.9\times10^{6}$	$5.2–7.8\times10^{6}$	$2.4–3.6\times10^{10}$
ATLAS equivalent	$3.5–5.2\times10^{13}$	$2.6–3.9\times10^{9}$	$5.2–7.8\times10^{9}$	$2.4–3.6\times10^{13}$

- The resources currently available are O(1000) off for full simulation for FSR
  - Might be ok for parametrized simulation
- Numbers should be multiplied by the number of detector variations and analysis
  - Although some optimisation might be possible

What next: S&C

- Investigate possibility to get more resources - e.g. spare cycles from WLCG - and be ready to use efficiently all what becomes available
  - See L. Valentini, iLCDirac
- Increase quality and efficiency of code
  - Long and expensive process, in general.
  - New faster simulation techniques promising (see A. Zaborowska talk)
    - Need to understand how to go beyond the full sim training statistics
- Investigate possibility selective/filtered simulations
  - Filters at generation level
  - Simulate only parts of relevance
- Facilitate interplay targeted full simulation and parametrized simulation
  - E.g. Automatic/optimal creation of Delphes configurations

What next: Physics Performance, ...

- Investigate (statistical) technologies to go beyond the rule of thumb

  MC Sample = Expected data sample

  reducing the number of events required
  - Could be useful also in perspective, when data will be there
- LHC is testing at similar statistics
  - Use this to identify processes requiring more attention

Final remarks

- Available computing resources are limited and do not allow full statistics studies
  - This is rather normal, given the investment they would represent
- Improvements in the code are always possible, but not such to change completely the picture
- MC implements our knowledge, should now what to expect
- We should try to use all that to identify possible criticalities where to use the available resources

Thanks!

From full sim to parametrized sim to phys perf

Dreaming bigger

- Common reconstruction tools between detector concepts?
  - Quite challenging
    - The optimal solution always depends on the specific features of the detector
  - Stating the obvious: write reconstruction algorithm as generic as possible
    - For simple cases: optimal solution for a given detector by tuning few parameters
    - For complex cases (e.g. Particle Flow): orchestration of modular tools that each detector implementation can arrange, tune or completely overwrite
- Ease (automatize?) the translation between detector performance evaluation from full sim and parametrized simulation
  - Allows us to sweep detector free parameters, probe their comprehensive impact on physics performance
- …