Cudacpp spin-off: a first study of MadGraph power consumption

Daniele Massaro
CERN, IT-FTI-PSE

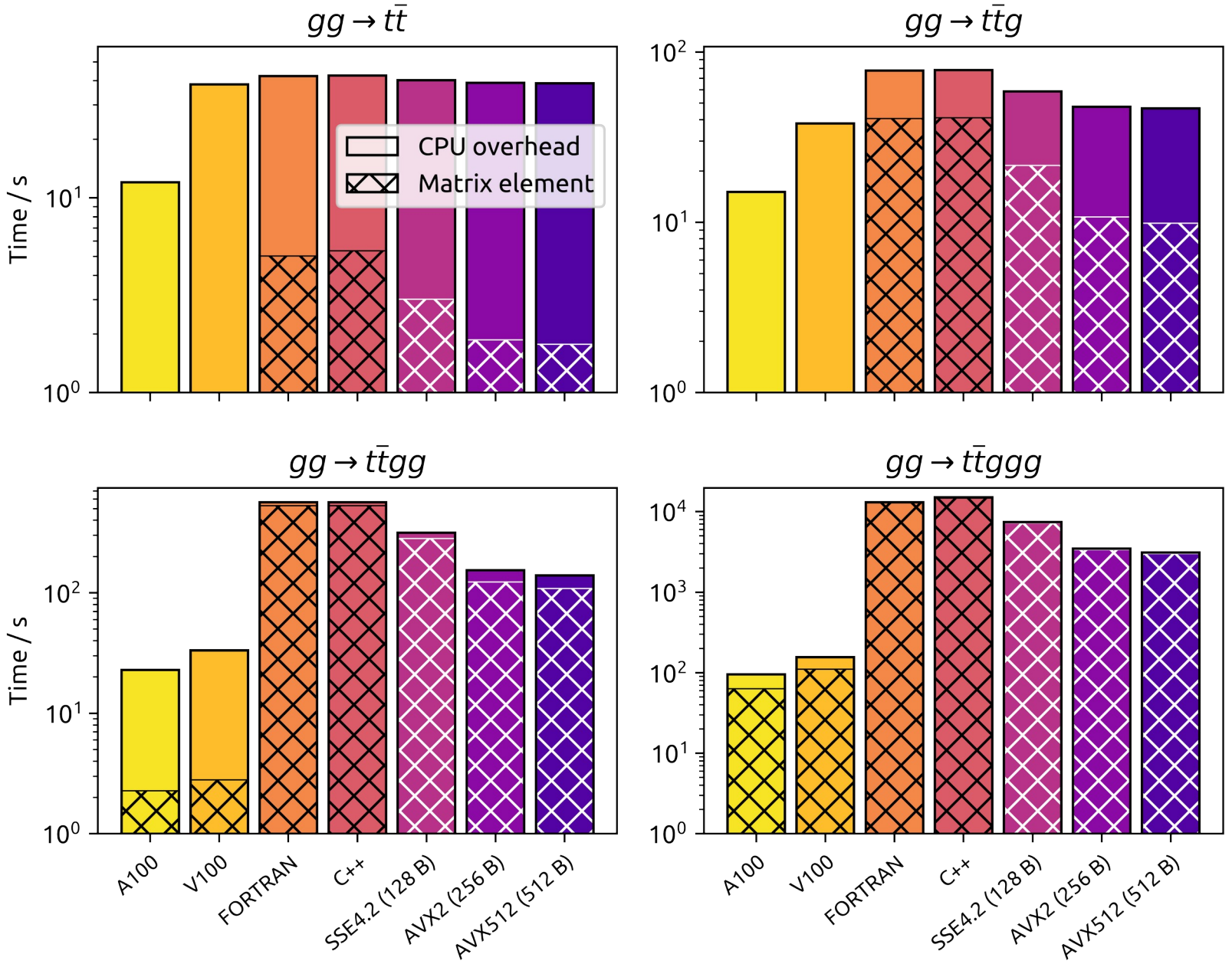
MadGraph5_aMC@NLO meeting 2025
3rd Feb 2025

Many thanks to Markus Schulz and Domenico Giordano for their precious suggestions and help to get performance metrics out of our machines.

Many thanks to Zenny Wettersten for the CO_2 numbers results.

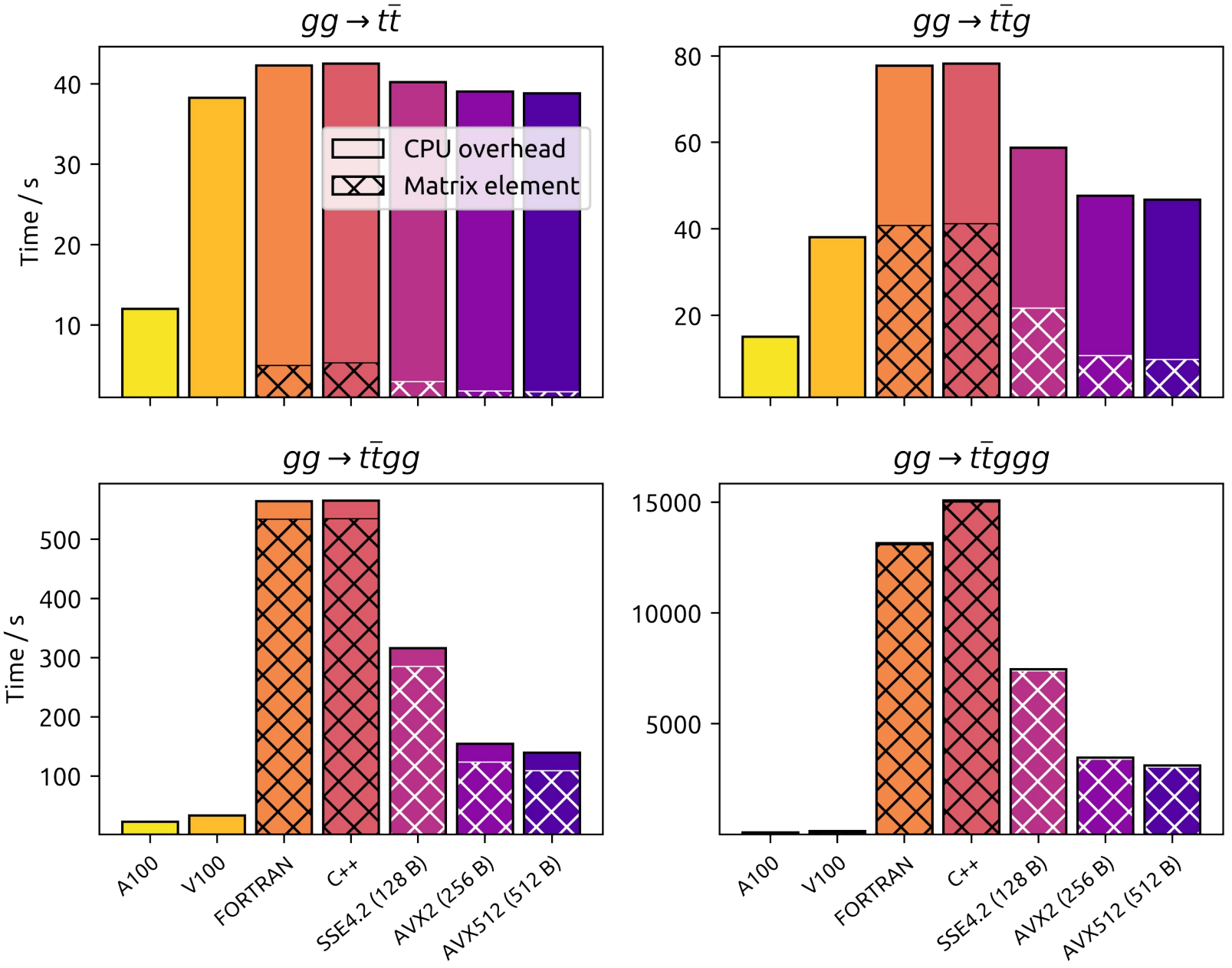
Performance comparison

Madevent single thread execution, 10^6 events



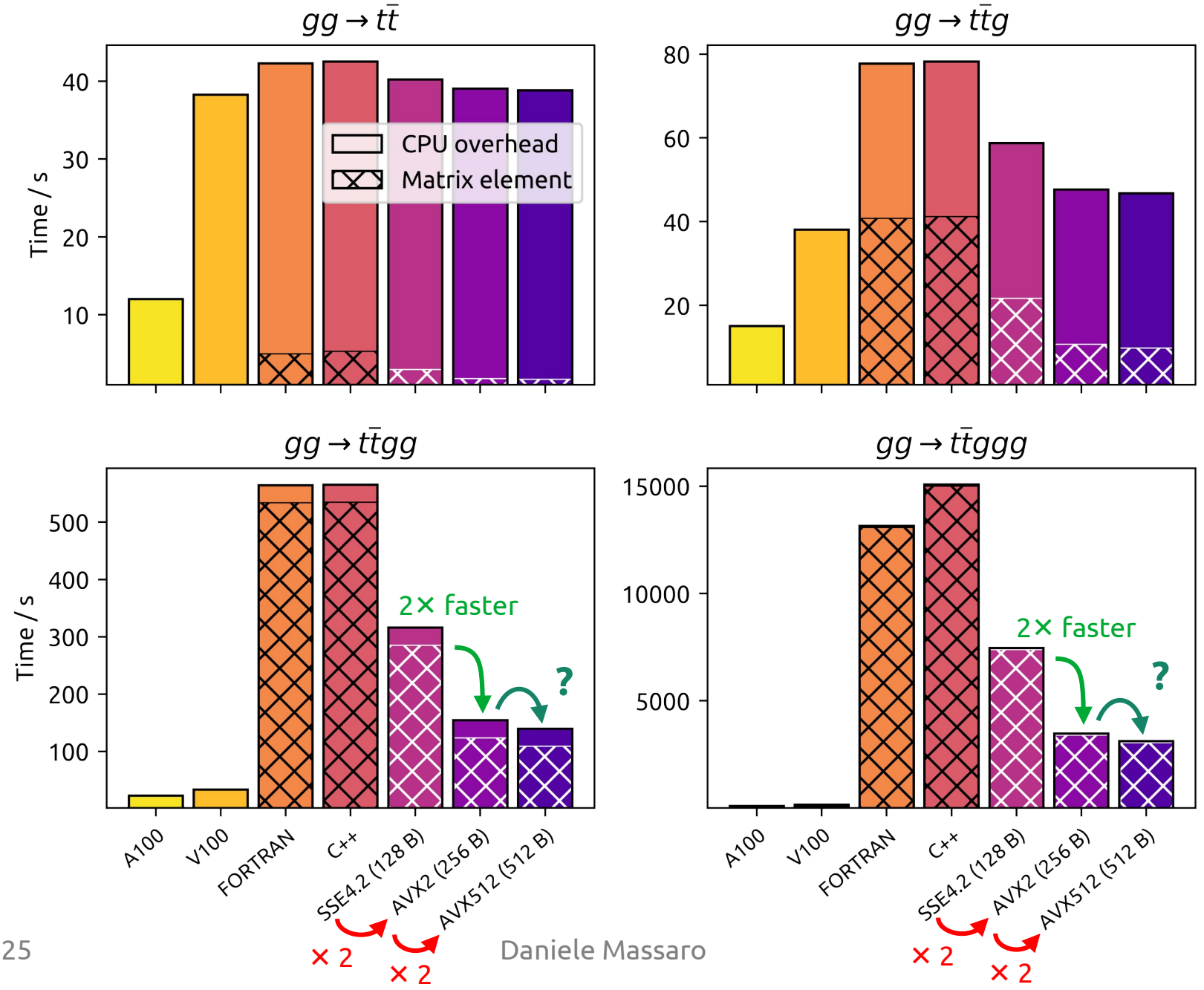
Performance comparison

Madevent single thread execution, 10^6 events

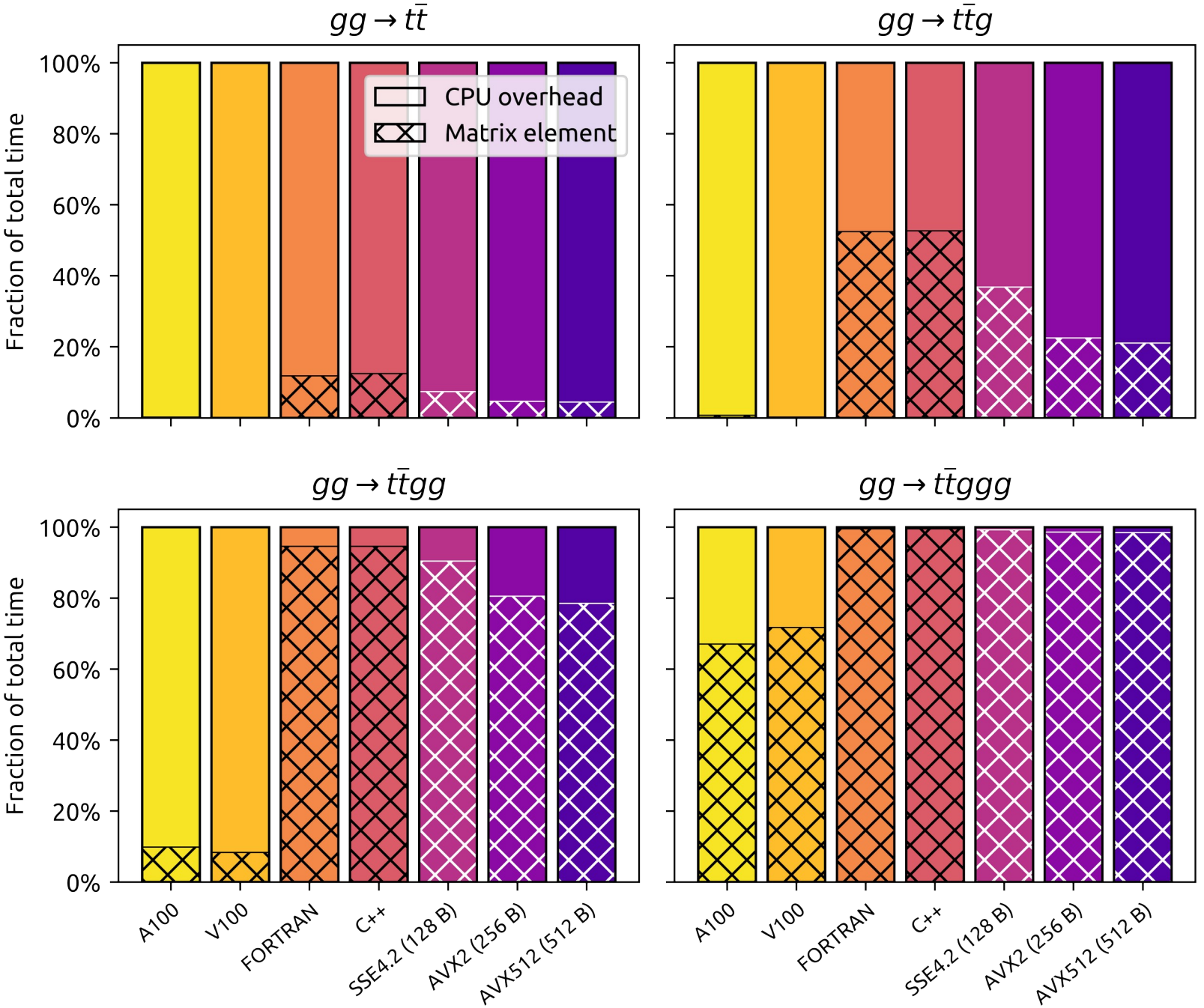


Performance comparison

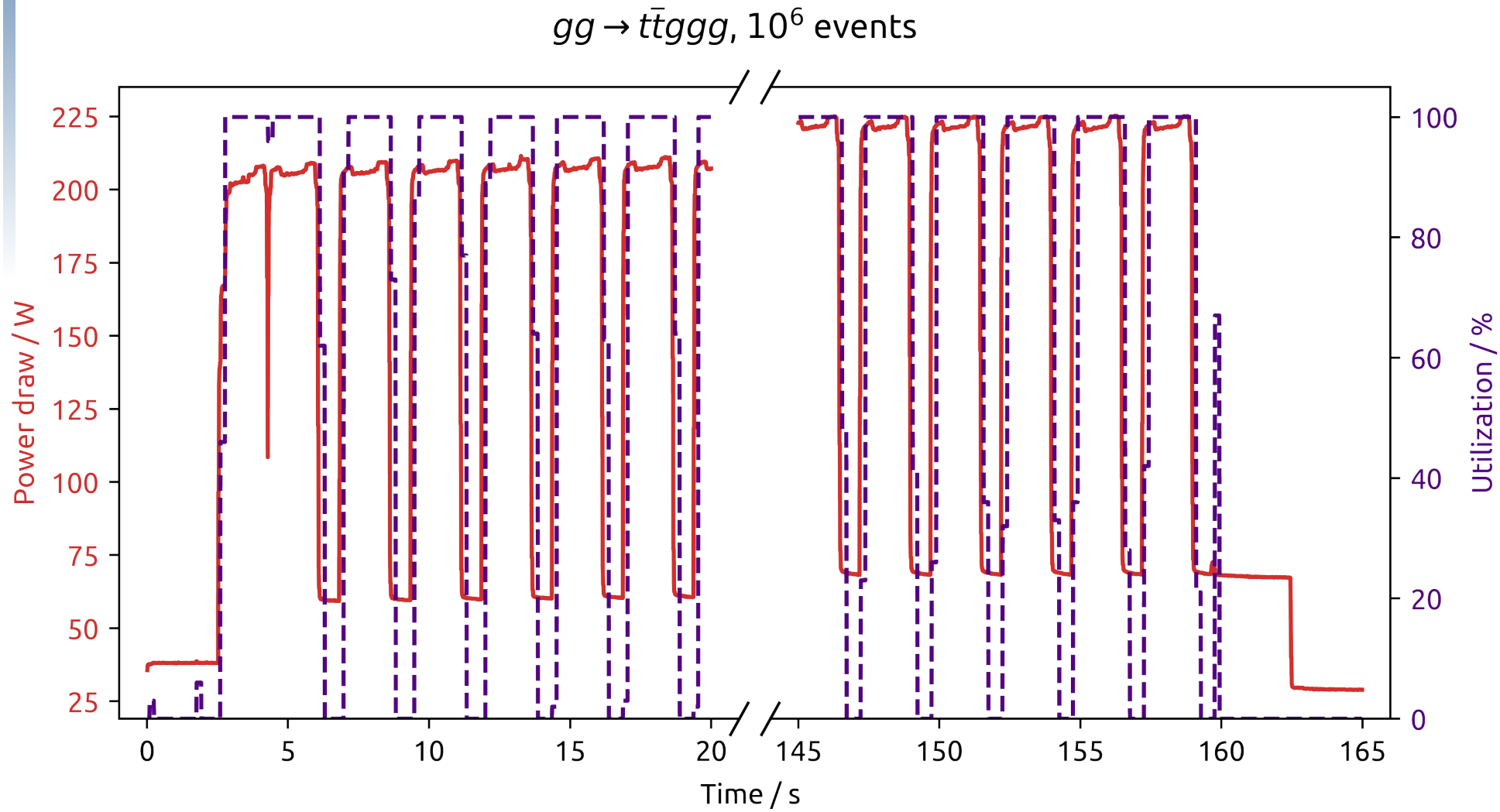
Madevent single thread execution, 10^6 events



Performance comparison

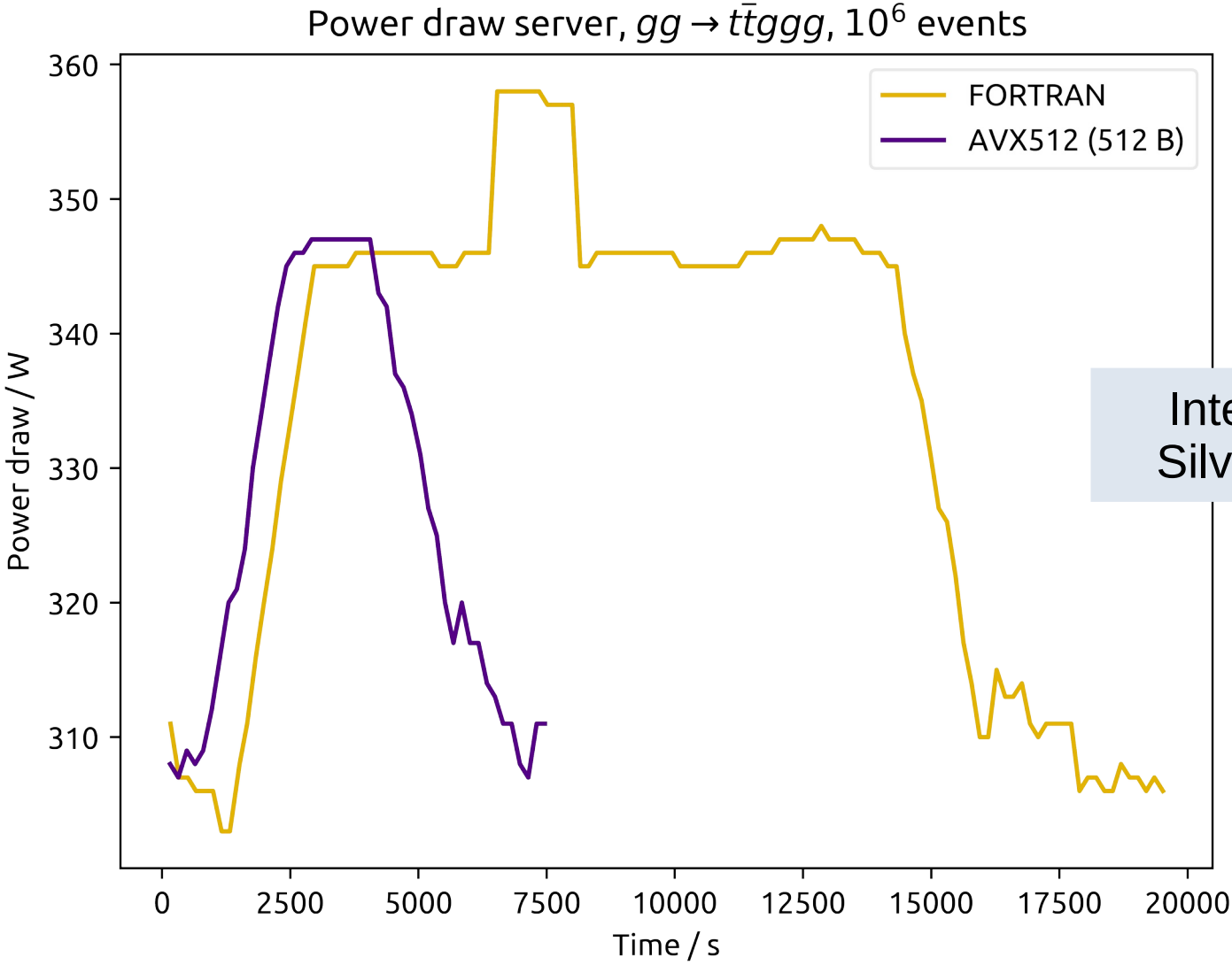


GPU (NVIDIA V100) only power draw



Power measurements obtained through `nvidia-smi`.

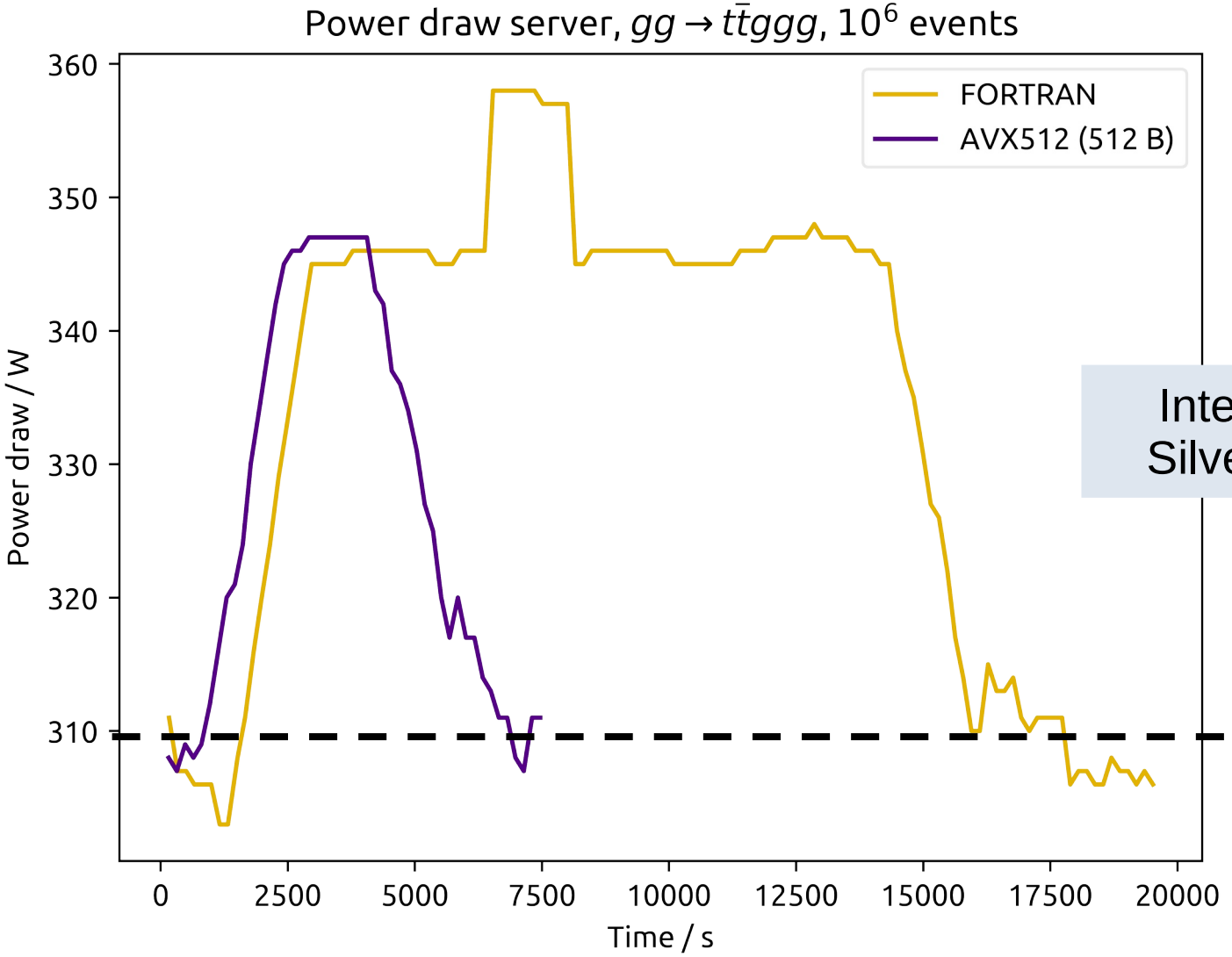
Server (CPU) power draw



*does include also normal server operations power draw.

Intel Xeon Silver 4216

Server (CPU) power draw



*does include also normal server operations power draw.

Intel Xeon Silver 4216

Server baseline power draw: 310 W.

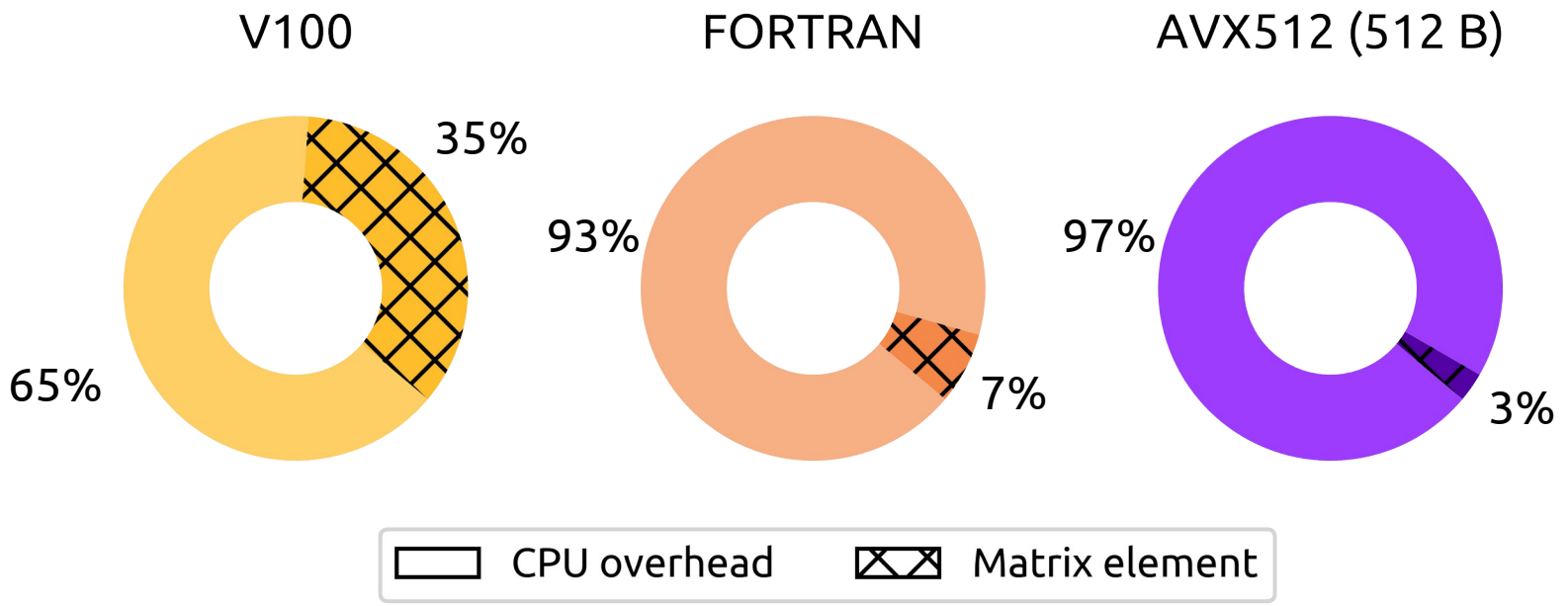
Energy consumption

- Server baseline power draw = 310 W
- 10^6 events.

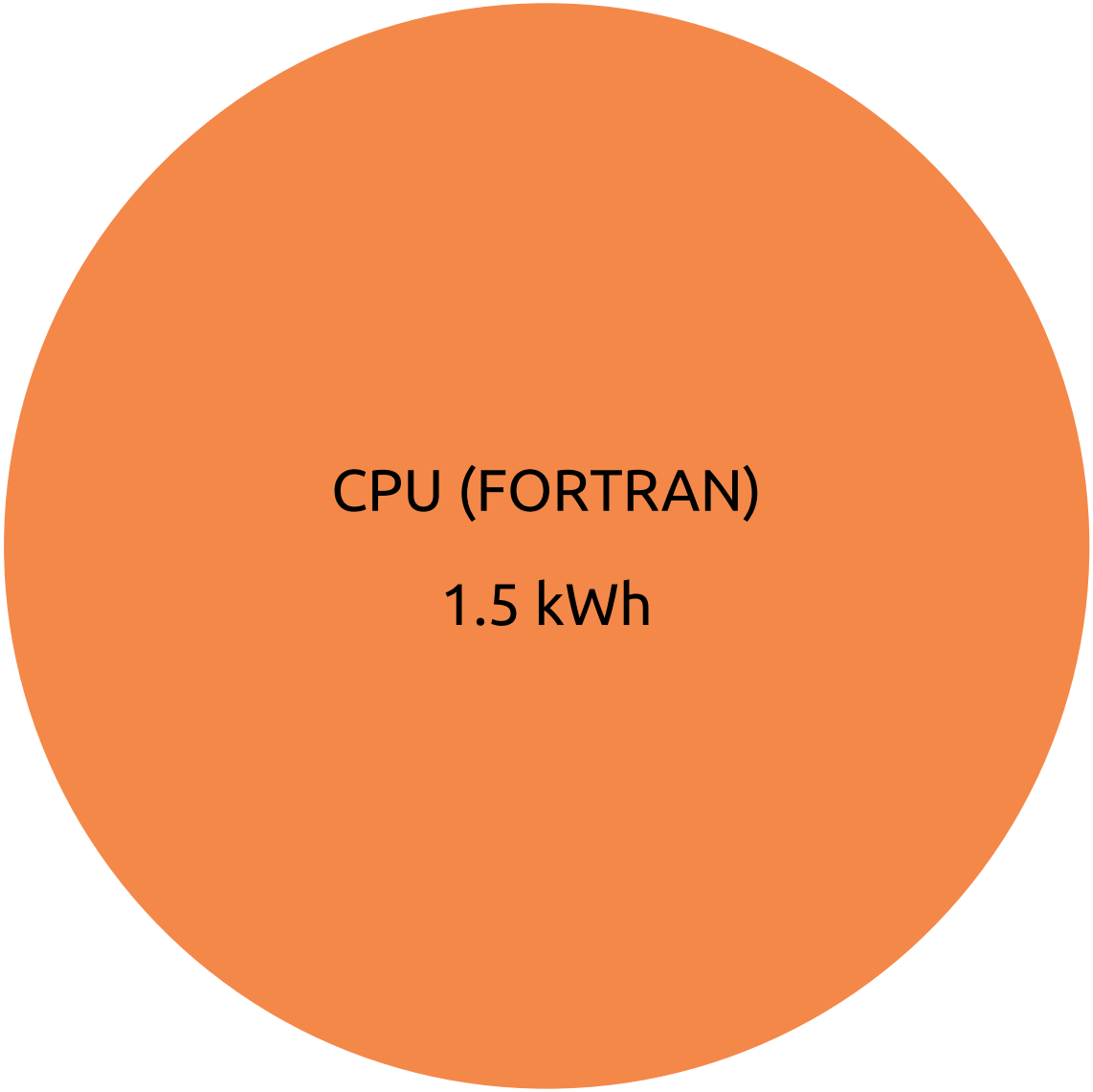
Backend	Net energy / kWh	Total “server” energy / kWh
GPU (V100)	0.0076	0.022
FORTRAN	0.11	1.5
AVX512	0.019	0.67

Energy consumption

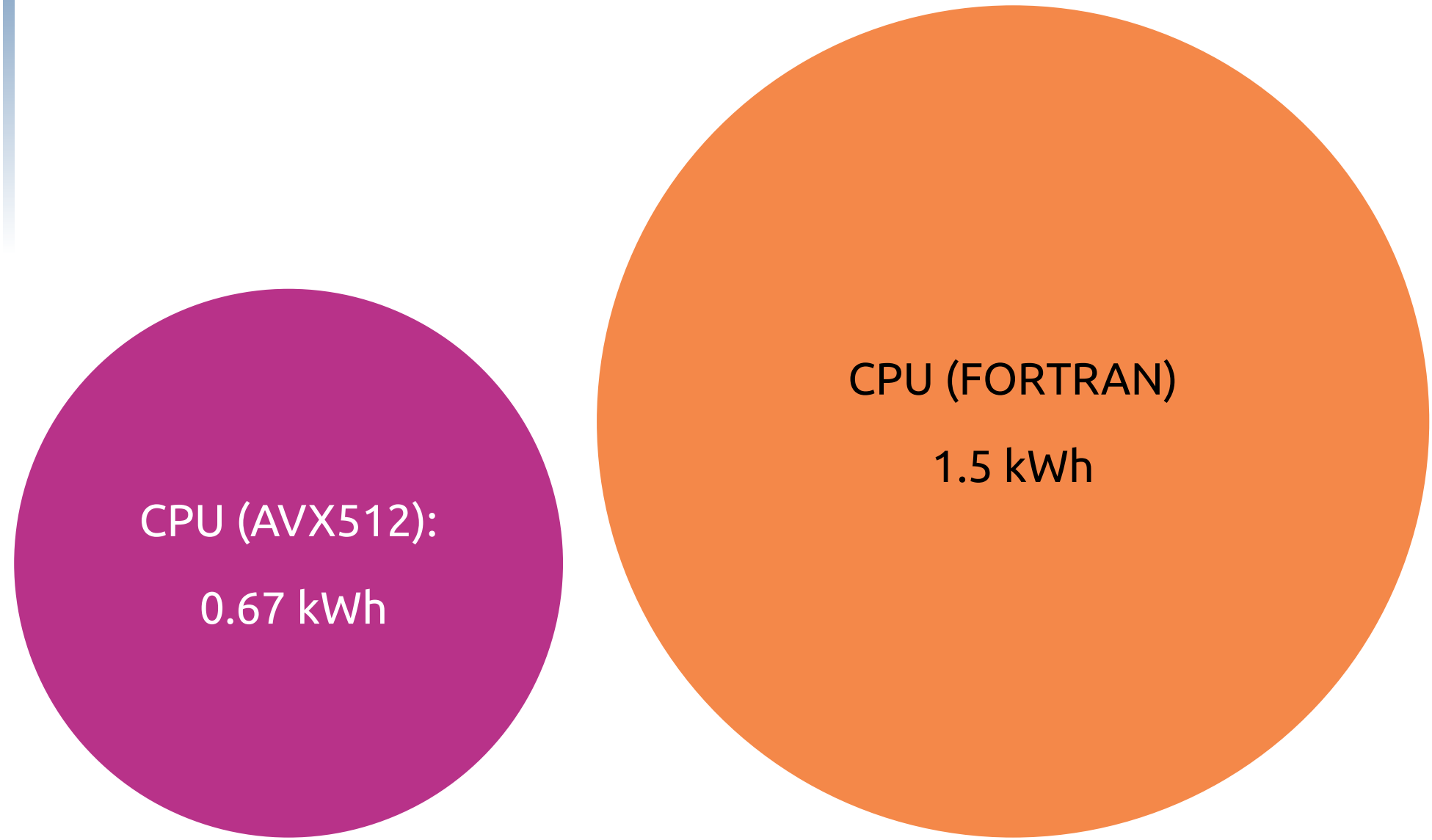
- Server baseline power draw = 310 W
- 10^6 events.



Energy consumption



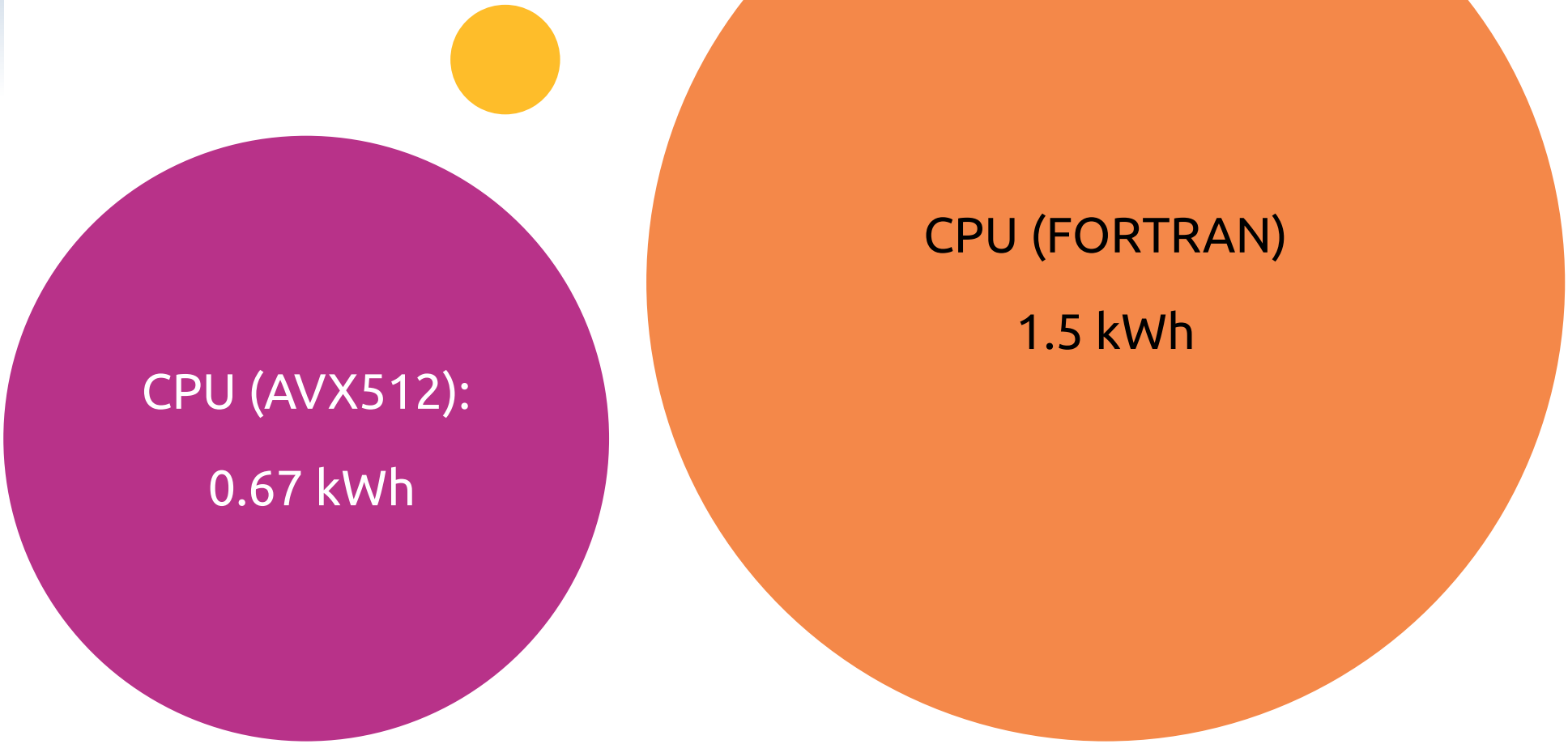
Energy consumption



Energy consumption

GPU (NVIDIA V100):

0.022 kWh



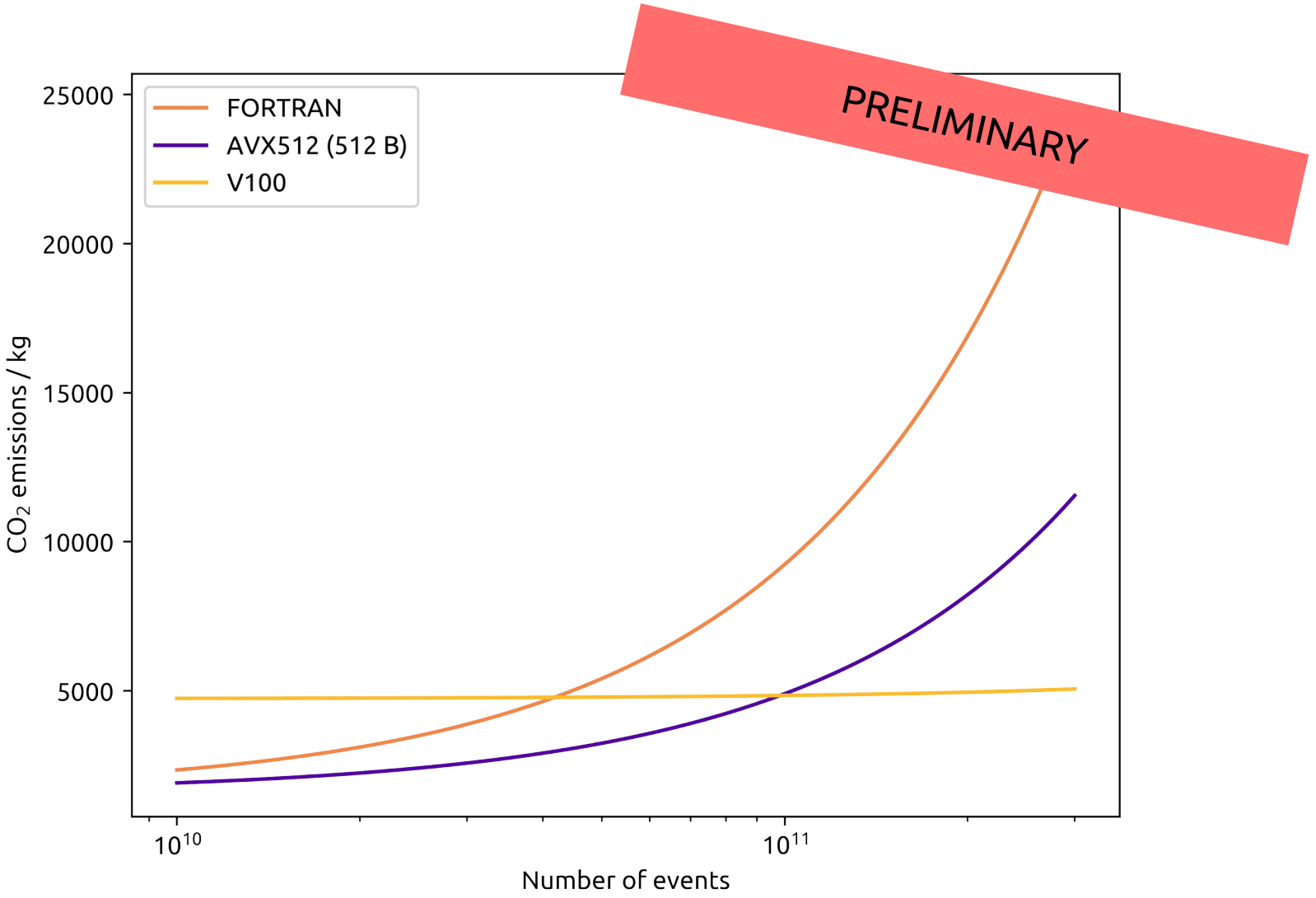
GPU vs CPU sustainability

- Embodied CO_2:
  - Assume lifetime of around 5 years.
  - Estimate to 0.02 g CO_2/s => 3155 kg.
- CO_2 emissions in France: 0.05 kg / kWh.

[https://www.eea.europa.eu/en/analysis/indicators/greenhouse-gas-emission-intensity-of-1]

How many events so that GPU and CPU CO_2 match?

GPU vs CPU sustainability



*assume the CO2 emissions are proportional the number of generated events.

Conclusions

- Main assumption: we calculated the compute as a function of the problem size…
- ...but reality is more complicated (you get limited amount of compute, limited access time, specific machines…).
- GPU:
  - decrease compute time => decrease total work;
  - but in the carbon there is a lot of carbon embedded.
- Next:
  - Further measurements on other machines (preferably baremetal).
  - Possibly test other hardware.
  - Refine CO2 numbers computation (e.g.: consider hardware lifetime).

Khazzoom-Brookes postulate

Energy efficiency improvements that, on the broadest considerations, are economically justified at the microlevel, lead to higher levels of energy consumption at the macrolevel.

- Modern version of the Jevons Paradox.

Khazzoom-Brookes postulate

Energy efficiency improvements that, on the broadest considerations, are economically justified at the microlevel, lead to higher levels of energy consumption at the macrolevel.

- Modern version of the Jevons Paradox.

[https://www.mondialframec.com/en/blog/gelato-on-the-road/]

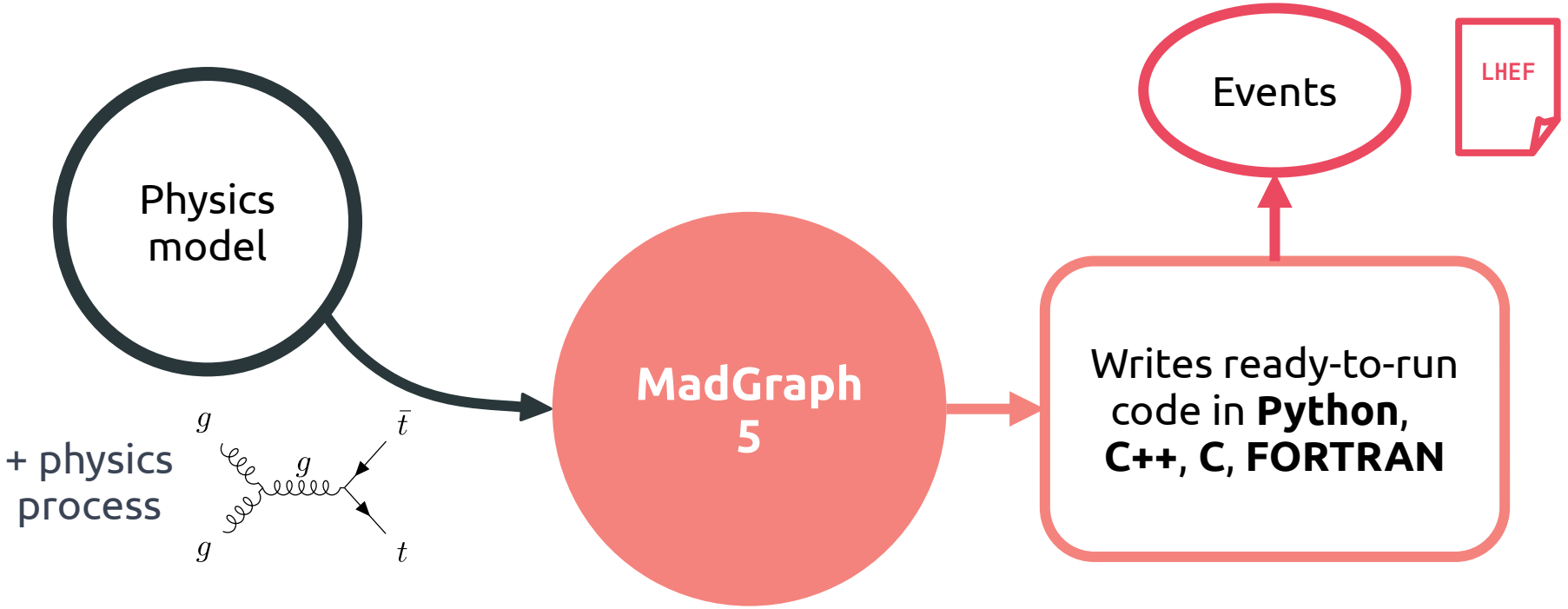
This icecream is so good and has only half the calories.
Well, then I'll take a triple scoop.

[https://youtu.be/6_LMkeclN7E?feature=shared]

Backup

[J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer, and T. Stelzer. JHEP 06 (2011), p. 128]
[J. Alwall et al. JHEP 07 (2014), p. 079]

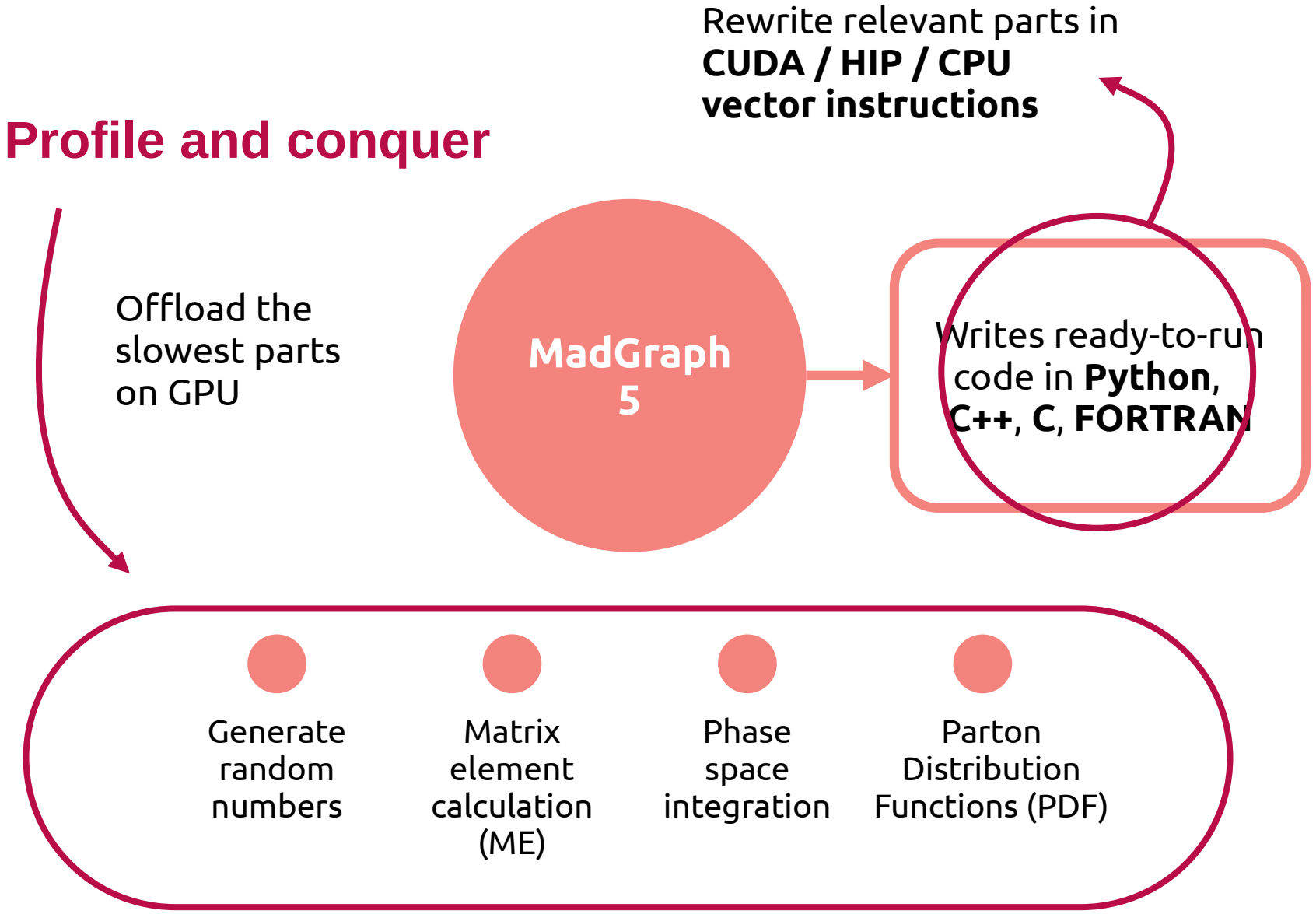
MadGraph 5



- Generate random numbers
- Matrix element calculation (ME)
- Phase space integration
- Parton Distribution Functions (PDF)

[J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer, and T. Stelzer. JHEP 06 (2011), p. 128]
[J. Alwall et al. JHEP 07 (2014), p. 079]

MadGraph 5

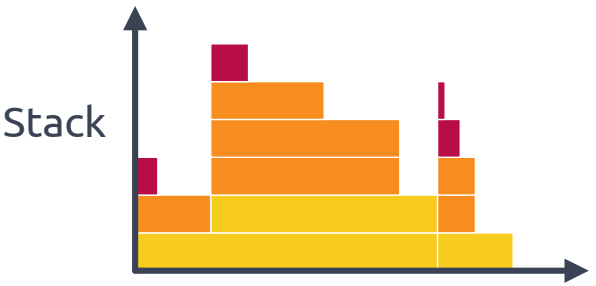


[J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer, and T. Stelzer. JHEP 06 (2011), p. 128]
[J. Alwall et al. JHEP 07 (2014), p. 079]

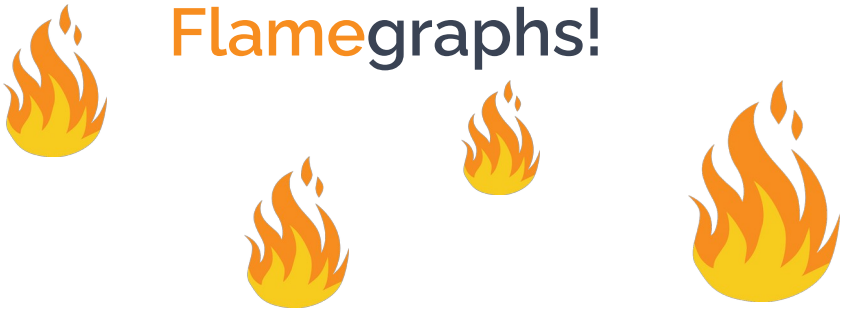
MadGraph 5

Profile and conquer

Offload the slowest parts on GPU



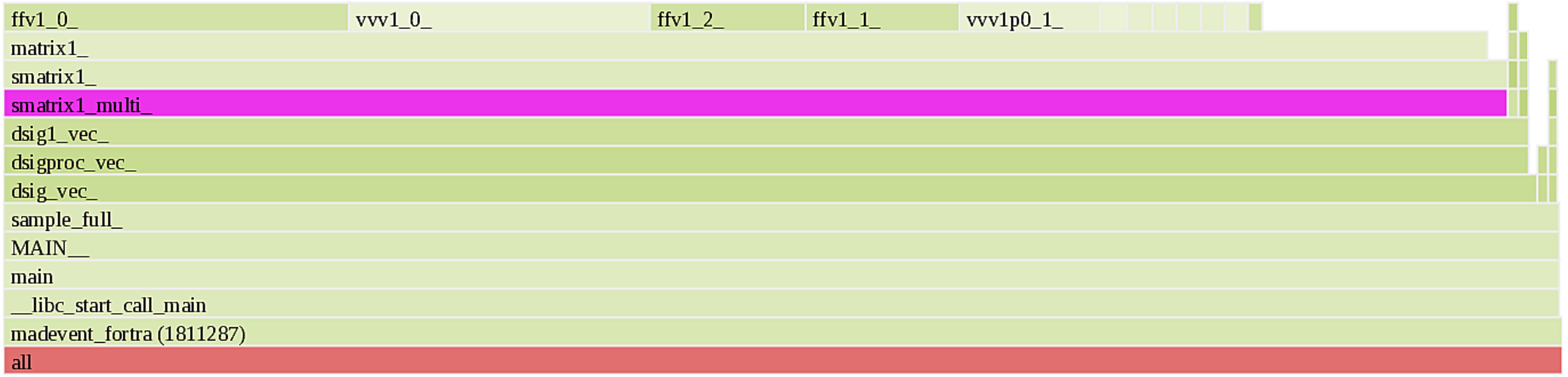
Flamegraphs!



- Generate random numbers
- Matrix element calculation (ME)
- Phase space integration
- Parton Distribution Functions (PDF)

$gg \to t\bar{t}gg$: FORTRAN

~ 97% running time



Generate random numbers

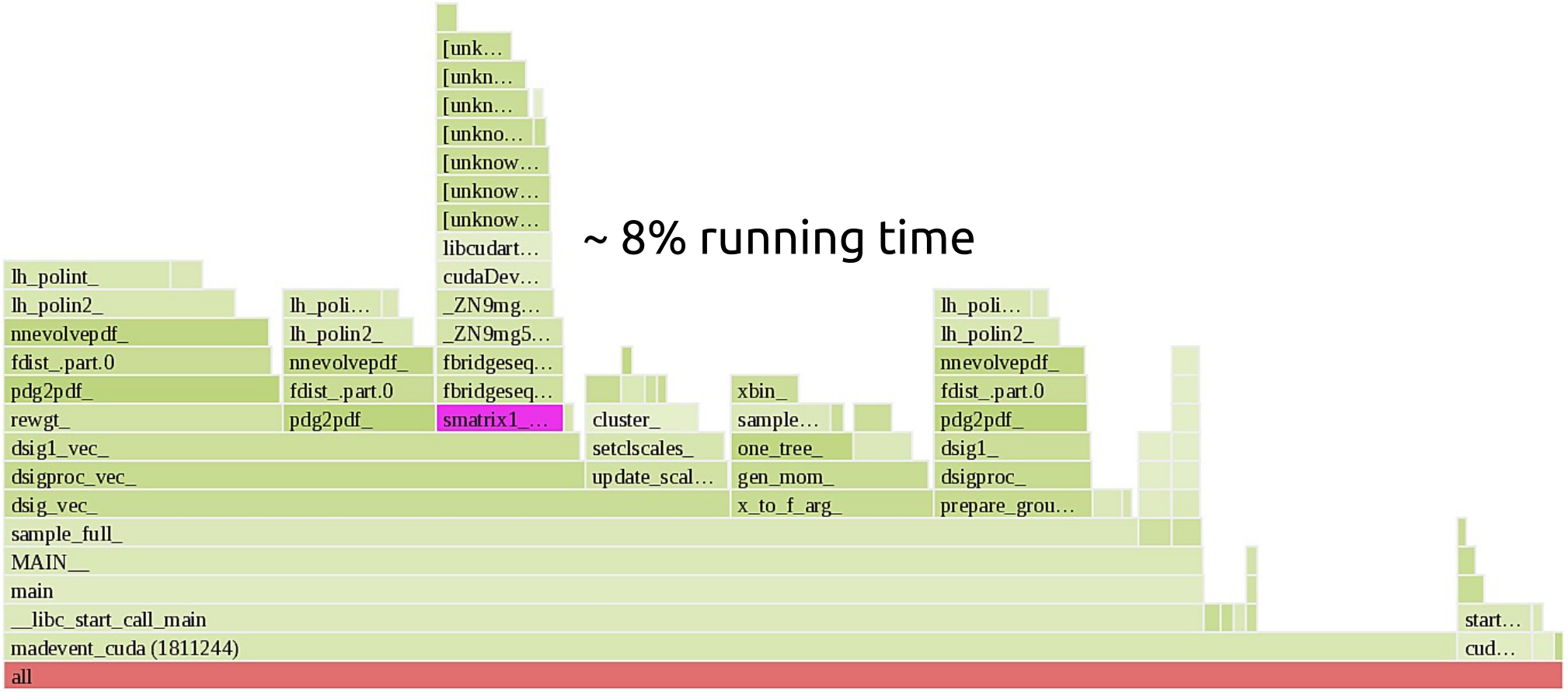
Matrix element calculation (ME)

Offload to GPU

Phase space integration

Parton Distribution Functions (PDF)

$gg \to t\bar{t}gg$: CUDA



~ 8% running time

[S. Hageboeck et al. EPJ Web Conf. 295 (2024), p. 11013]
[A. Valassi et al. ACAT 2022 (2023)]