Experiences of an HPC analyst in AI land

ISC workshop. June 26th, 2026
Jesus Labarta(jesus.labarta@bsc.es) , BSC

HORIZON-EUROHPC-JU-2023-COE

EuroHPC Joint Undertaking

1 January 2024– 31 December 2026

Grant Agreement No 101143931

POP: From HPC to AI ?

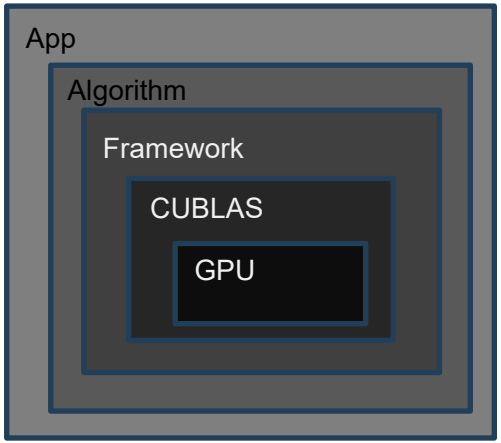
- HPC and AI ... speak a different language
  - Task, architecture, performance, … same words, DIFFERENT meaning
  - Mainly strong scaling vs. mainly weak scaling

- But still, we should be able to …
  - … build on HPC performance analysis tools and methodologies …
  - … towards seamless analysis of AI apps from very large to very small scale

How efficiently are we using our resources?

AI

- Black matryoshkas



- "no" visibility
  - What are we really doing?
  - How are we performing ?

vendor, …

- Just leave it to us !!
  - Our libraries, frameworks will do the BEST for you.
  - We do not offer mechanisms for you to take/steer our decision
- Why would you want explicit control of …
  - Kernel variant selected , #blocks, threads
  - In GEMMs, NCCLs, …

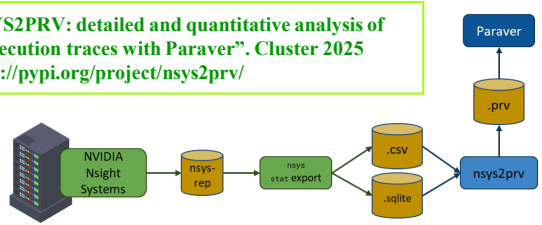
Assessments

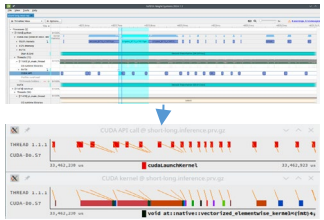
- Important to understand …
- … and may be optimize our use case?
  - A bit more informed decisions than exhaustive parameter exploration

nsys2prv

- Can we leverage Nsight Systems´s acquisition capability …
- and use Paraver to better squeeze the information in those traces ?
  - Scalability: dynamic range from very coarse to extremely fine grain
  - Analytics: timelines, histograms, correlations beyond profiles

M. Clascá et al. “NSYS2PRV: detailed and quantitative analysis of large-scale GPU execution traces with Paraver”. Cluster 2025
https://pypi.org/project/nsys2prv/





Towards microscopic insight …

… from single GPU to “large” clusters

A few tales …

A few tales

- Long, long time ago, …
- Common wisdom, real wisdom?
- Load imbalance in AI ?
- How important is communication ?
- What scares AI application engineers?
- You do not want that !!

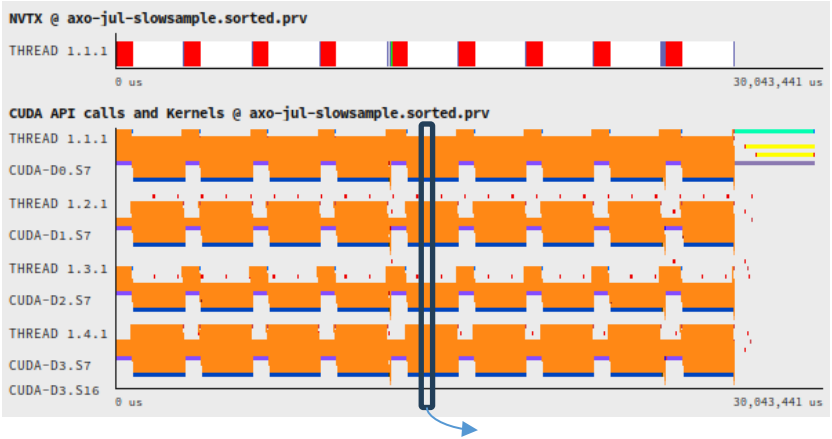
Long, long time ago

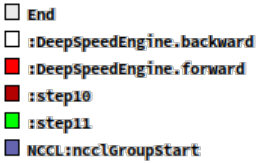
(24 months)

in a far, far …

My first Paraver view …

- … of an LLM training (fine tuning)
- @ 4 GPUs

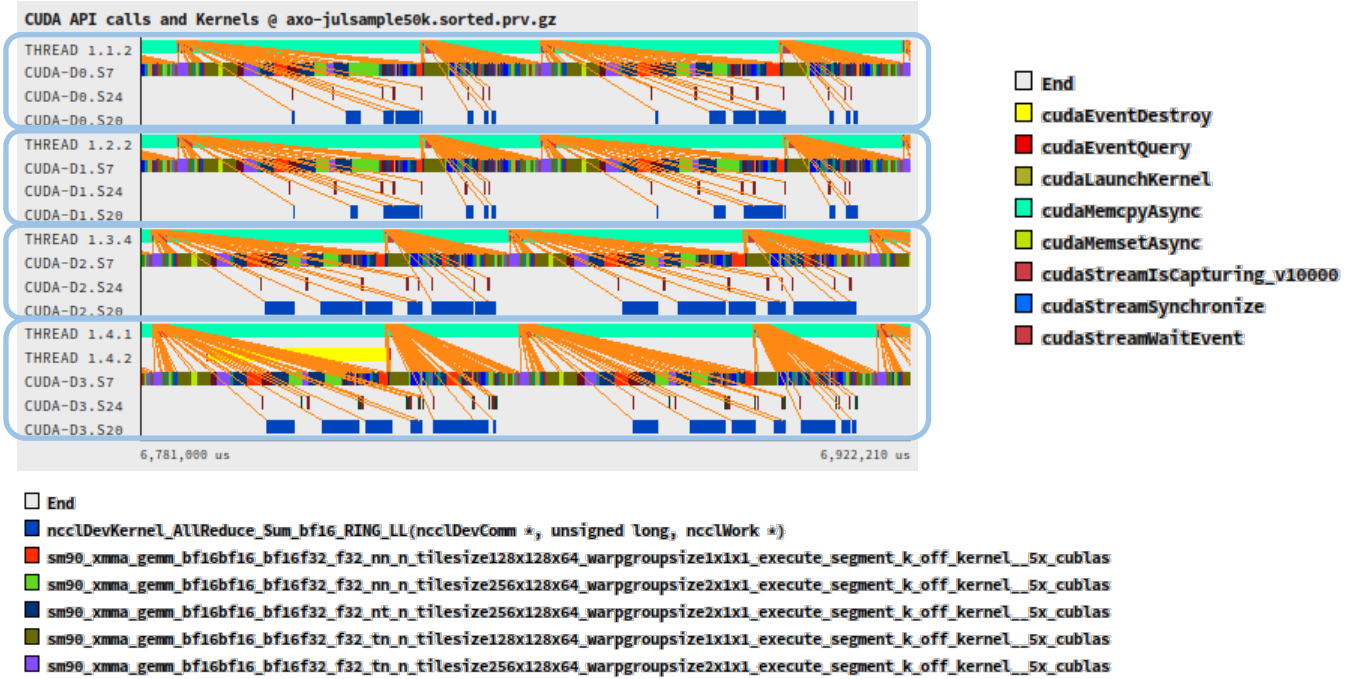




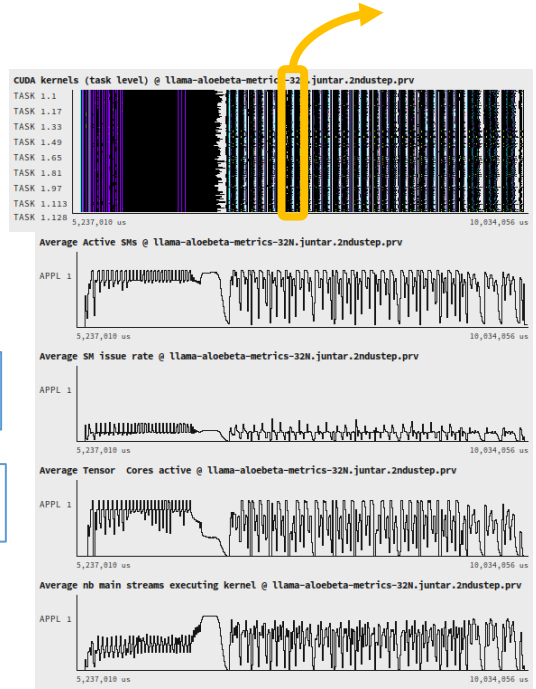
API calls and kernels

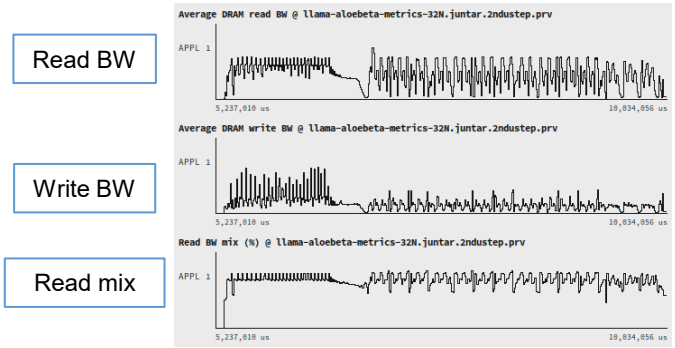
- Kernels, CUDA calls, Launch lines

GPUs & task based models?
Sparsity in performance data



@ 128 GPUs ?


Kernels
SM active
SM instr. Issue rate
Tensor core active
% active kernels
CUDA kernels (task level) @ llama-aloebeta-metrics-32N.juntar.2ndustep.prv
Average Active SMs @ llama-aloebeta-metrics-32N.juntar.2ndustep.prv
Average SM issue rate @ llama-aloebeta-metrics-32N.juntar.2ndustep.prv
Average Tensor Cores active @ llama-aloebeta-metrics-32N.juntar.2ndustep.prv
Average nb main streams executing kernel @ llama-aloebeta-metrics-32N.juntar.2ndustep.prv


Read BW
Write BW
Read mix
Average DRAM read BW @ llama-aloebeta-metrics-32N.juntar.2ndustep.prv
Average DRAM write BW @ llama-aloebeta-metrics-32N.juntar.2ndustep.prv
Read BW mix (%) @ llama-aloebeta-metrics-32N.juntar.2ndustep.prv

@ 128 GPUs ?

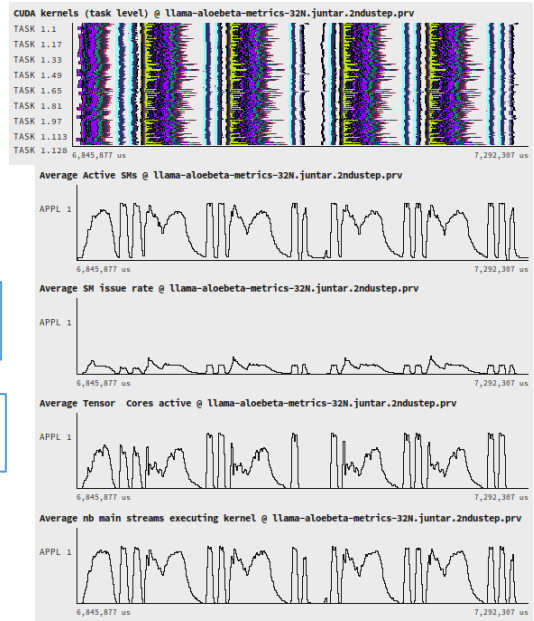
Kernels

SM active

SM instr. Issue rate

Tensor core active

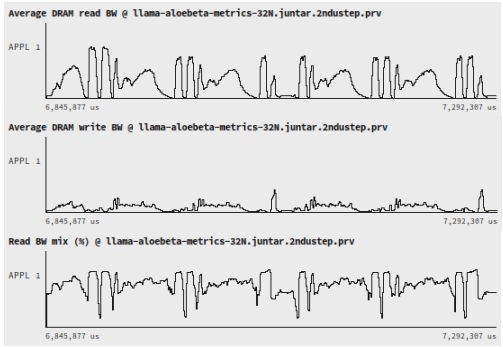
% active kernels



Read BW

Write BW

Read mix

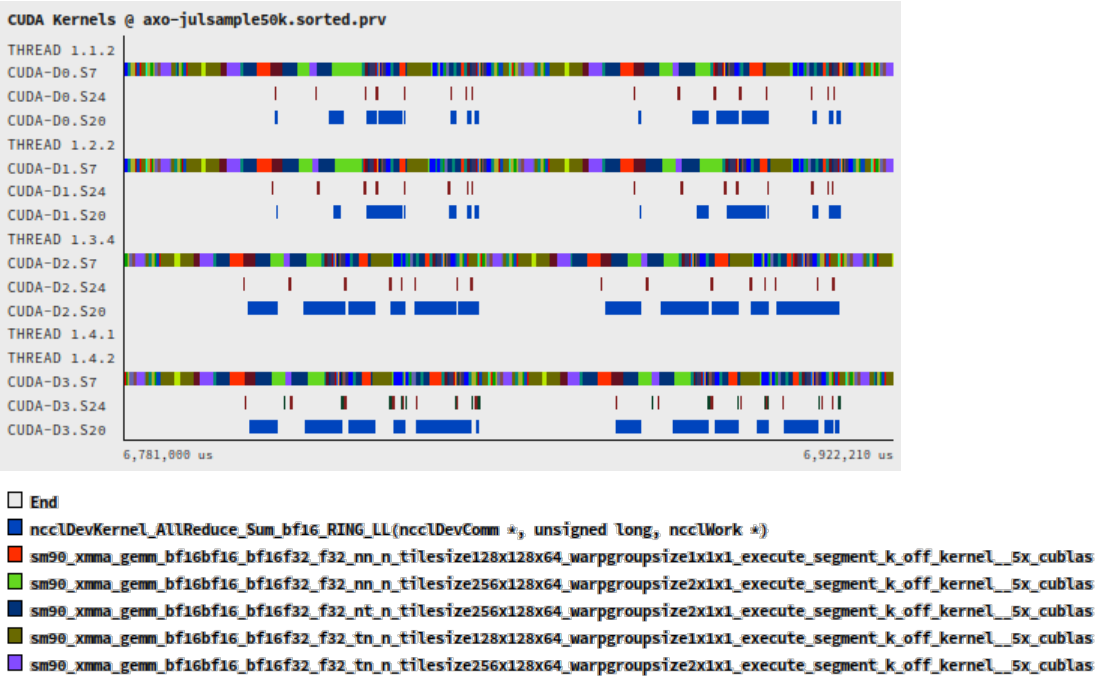


Lots of opportunities … who to “blame”?

Common wisdom … real wisdom?

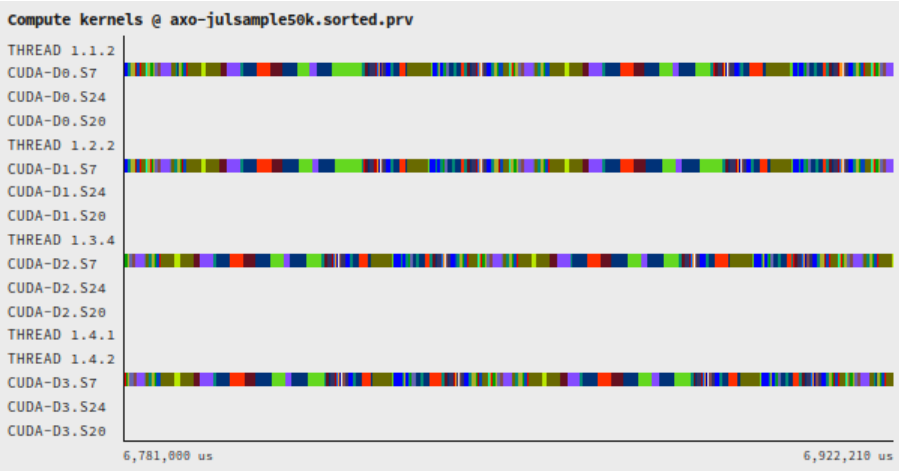
Common wisdom (I)

- “AI == communication bound” ?



Common wisdom (I)

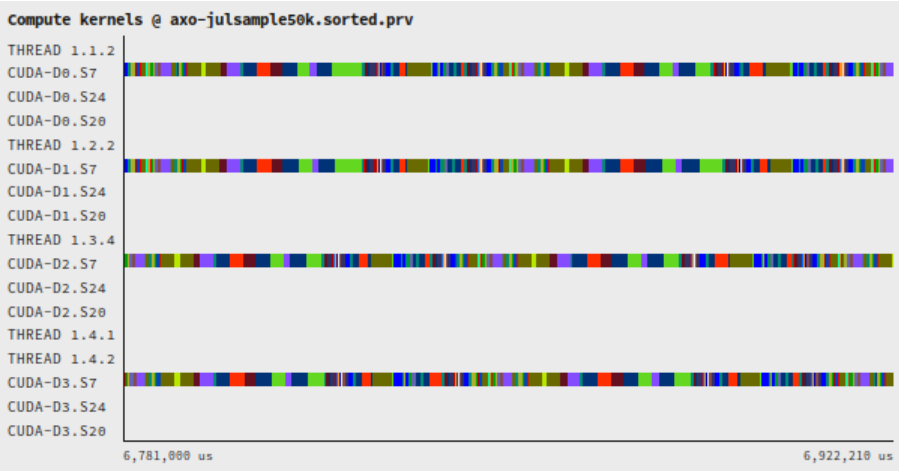
- "AI == communication bound"
  - … but this training case is certainly NOT communication bound



Idle %:	3.56

Common wisdom (II)

- “AI == MxM”



Idle %:	3.56

Common wisdom (II)

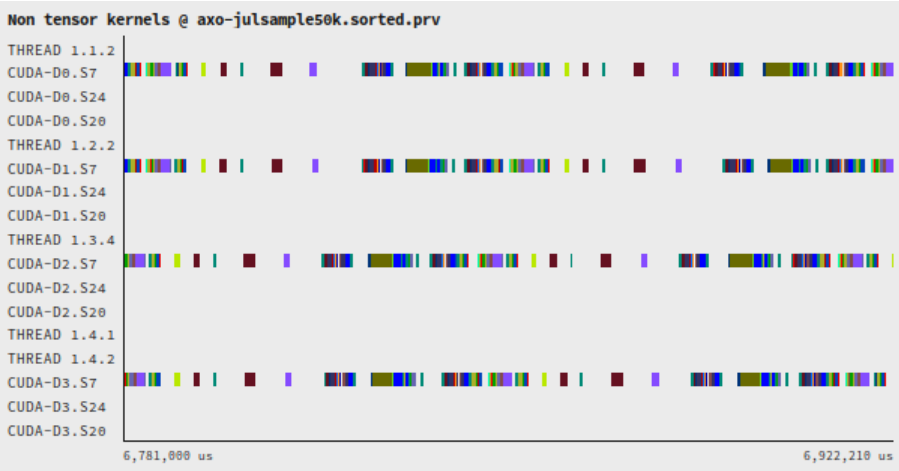
- "AI == MxM"
  - … at least in current platforms NOT the only/bottleneck

Idle %:	3.56
Gemm %:	49.95

sm90_xmma_gemm_bf16bf16_bf16f32_f32_nn_n_tilesize128x128x64_warpgroupsize1x1x1_execute_segment_k_off_kernel__5x_cublas
sm90_xmma_gemm_bf16bf16_bf16f32_f32_nn_n_tilesize256x128x64_warpgroupsize2x1x1_execute_segment_k_off_kernel__5x_cublas
sm90_xmma_gemm_bf16bf16_bf16f32_f32_nt_n_tilesize256x128x64_warpgroupsize2x1x1_execute_segment_k_off_kernel__5x_cublas
sm90_xmma_gemm_bf16bf16_bf16f32_f32_tn_n_tilesize128x128x64_warpgroupsize1x1x1_execute_segment_k_off_kernel__5x_cublas
sm90_xmma_gemm_bf16bf16_bf16f32_f32_tn_n_tilesize256x128x64_warpgroupsize2x1x1_execute_segment_k_off_kernel__5x_cublas

Common wisdom (II)

- “AI == MxM”
  - … at least in current platforms NOT the only/bottleneck

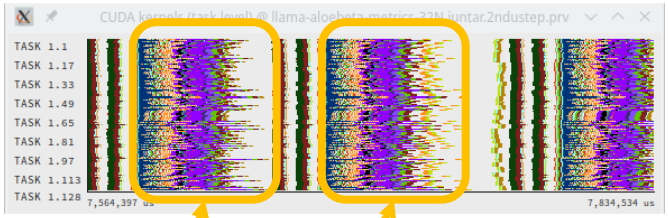


Idle %:	3.56
Gemm %:	49.95
Non Gemm %:	46.49

Load imbalance in AI?

@ 128 GPUs ?

- Kernels and duration

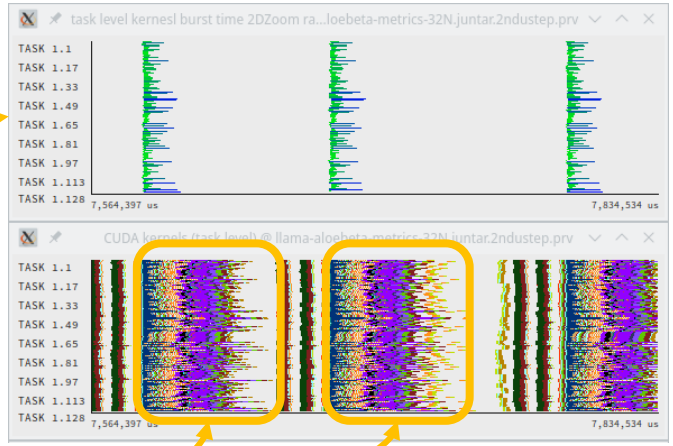


Load imbalance !!!
Repetitive pattern

Load balance:
77%

@ 128 GPUs ?

- Kernels and duration



flash_bwd_dq... kernel duration
Repetitive pattern

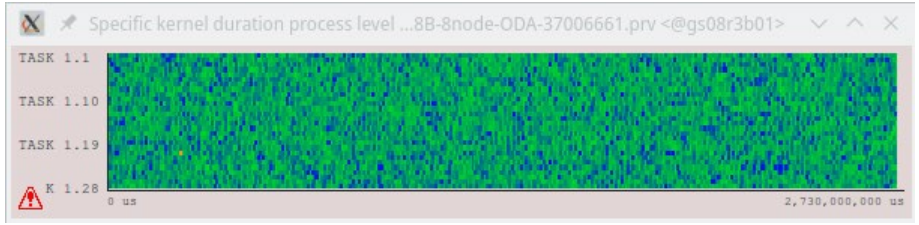
Load imbalance !!!
Repetitive pattern

Packing impact along time

- flash_bwd_… kernel duration

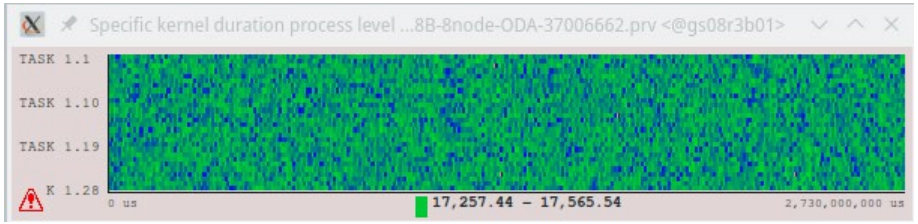
Sequence length 32K; 360 steps; 15 GB

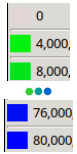
Sequential



Stalls: 1.12%

MP default





0

flash_bwd duration

2730 s

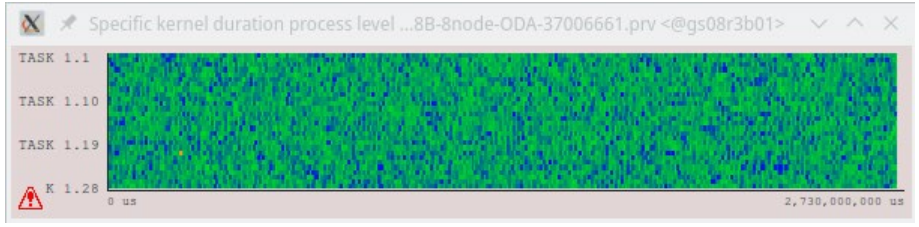
"Cost-Aware Packing Pipelines for Load Balanced LLM Fine-Tuning". Submitted

Packing impact along time

- flash_bwd_… kernel duration

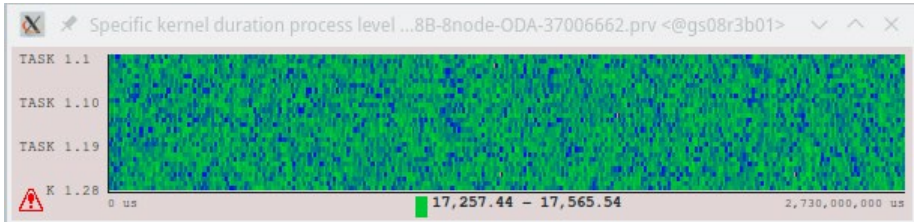
Sequence length 32K; 360 steps; 15 GB

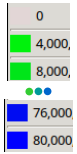
Sequential



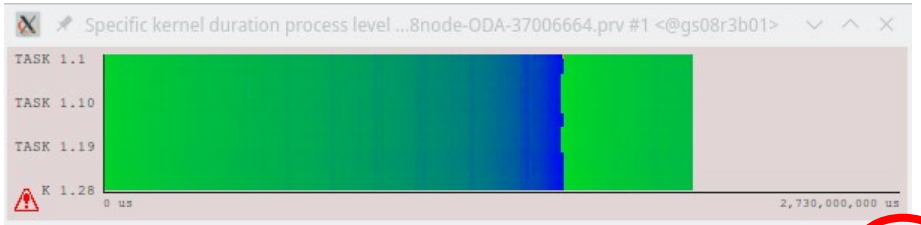
Stalls: 1.12%

MP default





Cycle MP Sort



Stalls: 2.56%

0

flash_bwd duration

2730 s

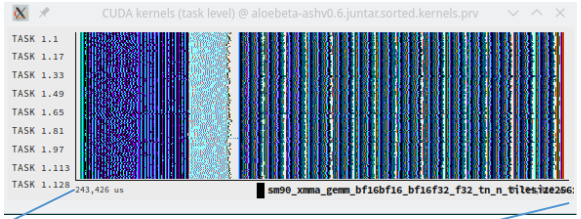
"Cost-Aware Packing Pipelines for Load Balanced LLM Fine-Tuning". Submitted

How important is “communication”?

To overlap or not to overlap

- Different microscopic behaviors … similar macroscopic duration !!!!

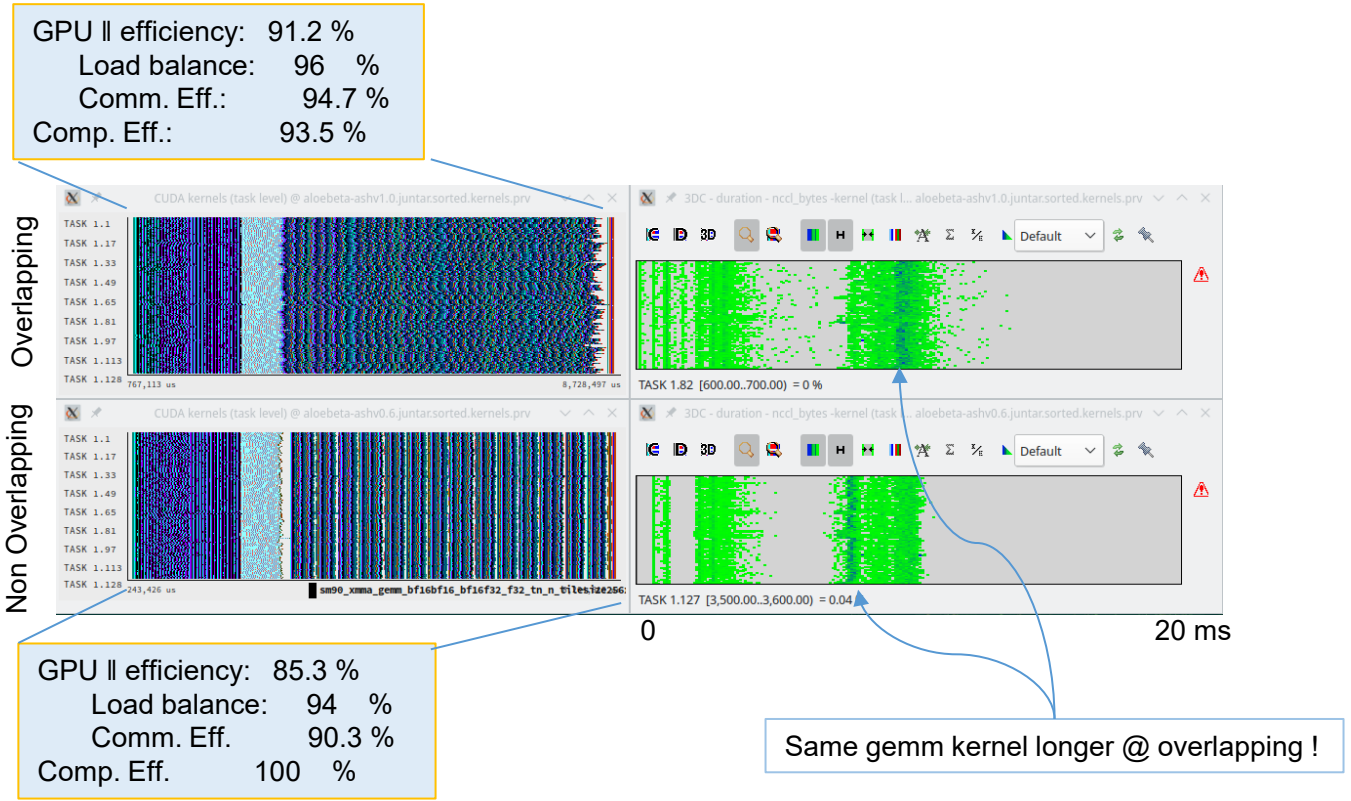
Non Overlapping



GPU ‖ efficiency: 85.3 %
Load balance: 94 %
Comm. Eff. 90.3 %
Comp. Eff. 100 %

To overlap or not to overlap

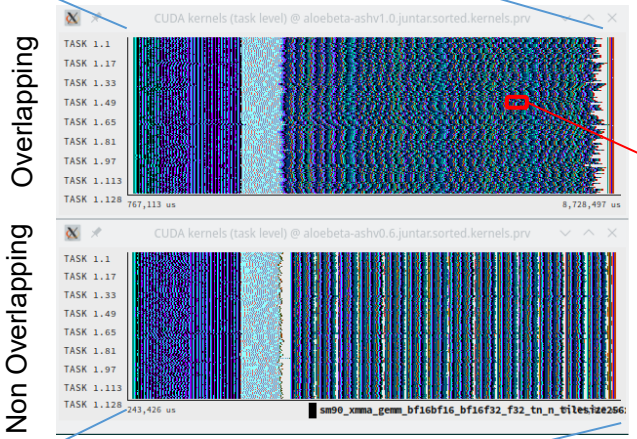
- Different microscopic behaviors … similar macroscopic duration !!!!



To overlap or not to overlap

- Different microscopic behaviors … similar macroscopic duration !!!!

GPU ‖ efficiency: 91.2 %
Load balance: 96 %
Comm. Eff.: 94.7 %
Comp. Eff.: 93.5 %

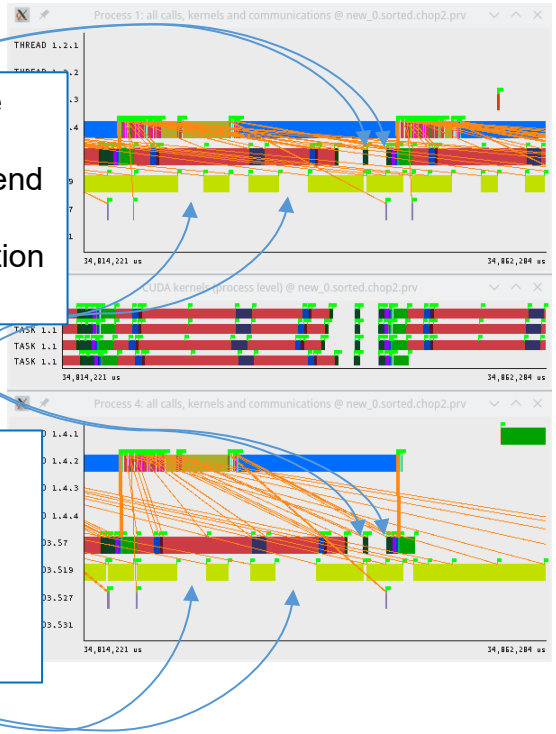


GPU ‖ efficiency: 85.3 %
Load balance: 94 %
Comm. Eff. 90.3 %
Comp. Eff. 100 %

Comp – Comm interferences ?

- Call it communication …
- … mean SM scheduling

Tight zoom in fwd



Actual Dependences

Compute kernels delayed till end of communication kernels

Resource contention !!!!

That may or may not move actual data transfer into the critical path

NCCL kernels delayed till “end” of compute kernels

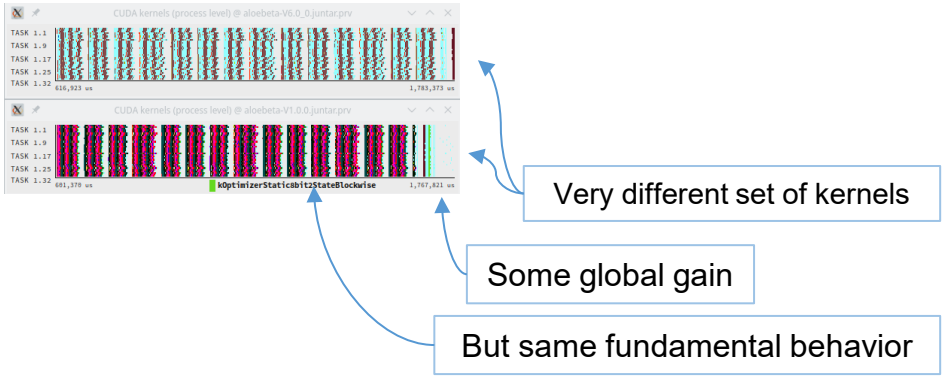
To conclude …

Life in AI times

“My dear, here we must run as fast as we can, just to stay in place.
And if you wish to go anywhere you must run twice as fast as that.”

Lewis Carroll. Through the Looking-Glass

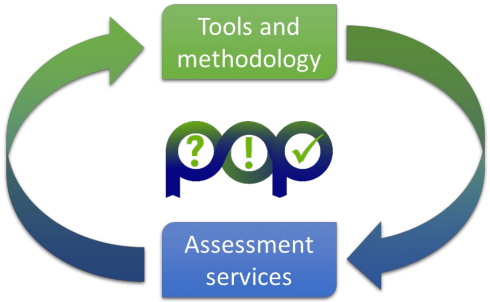
From when you say is the framework version?
Three months ago? … ☹ … update !!



But …
…we can leverage HPC Performance analysis tools, methodology and background to get external assessment of the behavior of AI apps !

POP

- Performance optimization and productivity COE



FREE Services provided by the CoE

- **Parallel Application Performance Assessment**
  - Primary service
  - Initial analysis measuring a range of performance metrics to assess quality of performance and identify the issues affecting performance (at customer site)
  - If needed, undertakes further performance evaluations to identify the root causes of the issues found and qualify and quantify approaches to address them (recommendations)
- **Second Level Services**
  - Second level services may follow after conclusion of an initial performance assessment:
    - **Proof-of-concept**: explore the potential benefit of proposed optimisations by applying them to selected regions of the applications
    - **Correctne… …y**: evaluate the correctness of hybrid MPI + OpenMP applications
    - **Ener… …udy**: investigate improvements of energy consumption or efficiency
    - …ng consultancy for customers that choose to implement proposed …own

…our experts and customer!

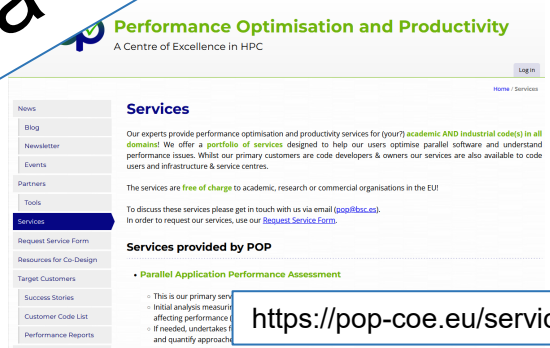
10

HORIZON-EUROHPC-JU-2023-…

Grant Agreem… …743931

1 January 2024– 3… December 2026

Give POP a try



https://pop-coe.eu/services

Performance Optimisation and Productivity 3

A Centre of Excellence in HPC

Contact:

- https://www.pop-coe.eu
- pop@bsc.es
- @POP_HPC
- youtube.com/POPHPC

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101143931. The JU receives support from the European Union's Horizon Europe research and innovation programme and Spain, Germany, France, Portugal and the Czech Republic.