

Performance Evaluation of LLM workloads on Novel Accelerators

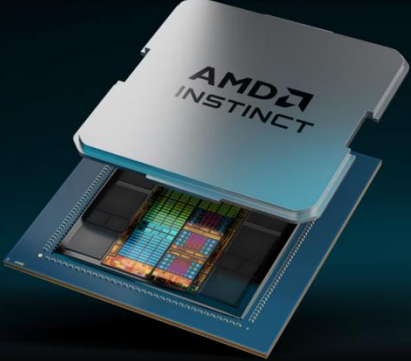
Lancelot Normand
McGill University

What's new?

- Systematic collection of data from finetuning experiments and traces across multiple state-of-the-art accelerators.
- Data reported to Weights and Biases
- Trace produced in CTF or Nsys.

Novel Accelerators

- AMD MI300a Unified memory
- GH200 (All in one chip, not unified)




AMD Instinct MI300A APUs

AMD Instinct MI300A accelerated processing units (APUs) combine the power of AMD Instinct accelerators and AMD EPYC™ processors with shared memory to enable enhanced efficiency, flexibility, and programmability. They are designed to accelerate the convergence of AI and HPC, helping advance research and propel new discoveries.

[View Specs >](#)

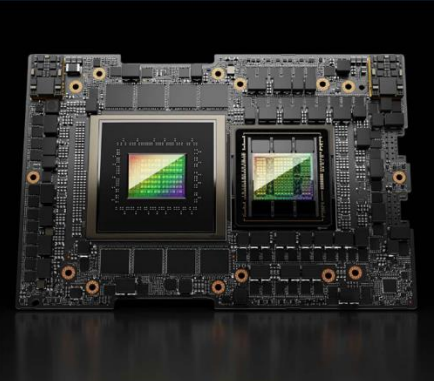
228 CUs	24	128 GB	5.3 TB/s
228 GPU Compute Units	24 "Zen 4" x86 CPU Cores	128 GB Unified HBM3 Memory	5.3 TB/s Peak Theoretical Memory Bandwidth

Datasheet



NVIDIA GH200 Grace Hopper Superchip

The breakthrough processor for large-scale AI and high-performance computing (HPC) applications.



Some challenges with working with new Acc.

- Libraries might not be optimized.
- System must be properly configured to fully take advantage of features such as unified memory.
- They may appear to underperform because of bad config of library support.

Accelerators and Models used

- AMD MI300A
- AMD MI210
- NVIDIA GH200
- NVIDIA H100
- Llama 3
- OPT (Open source of GPT 3)
- Bloom

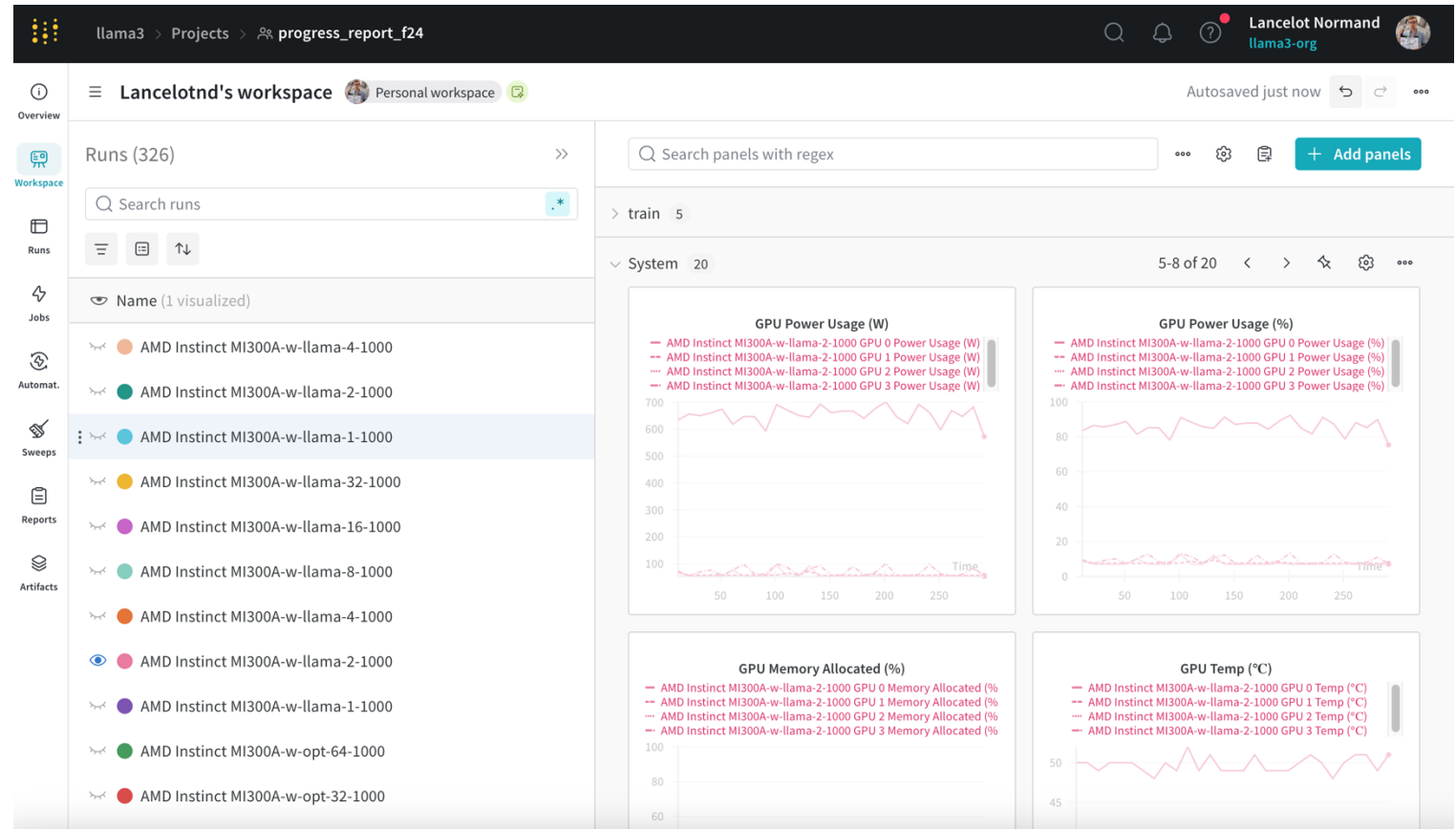
How we analyze the data?

 **Weights & Biases**



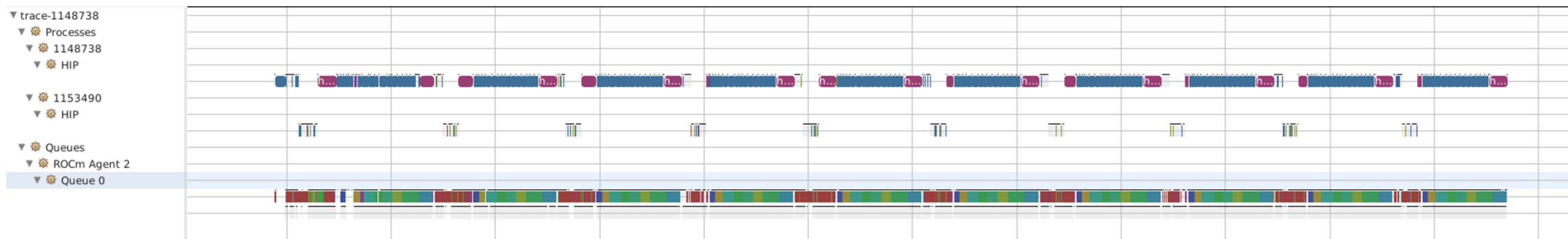
Using Wandb

- Can be done in user space
- Results are nicely recorded
- Mostly profiling data

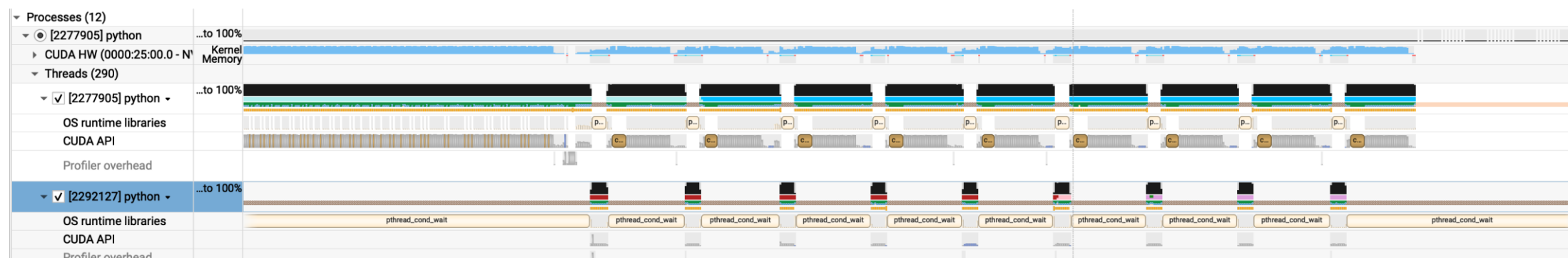


Finetuning for 10 iterations

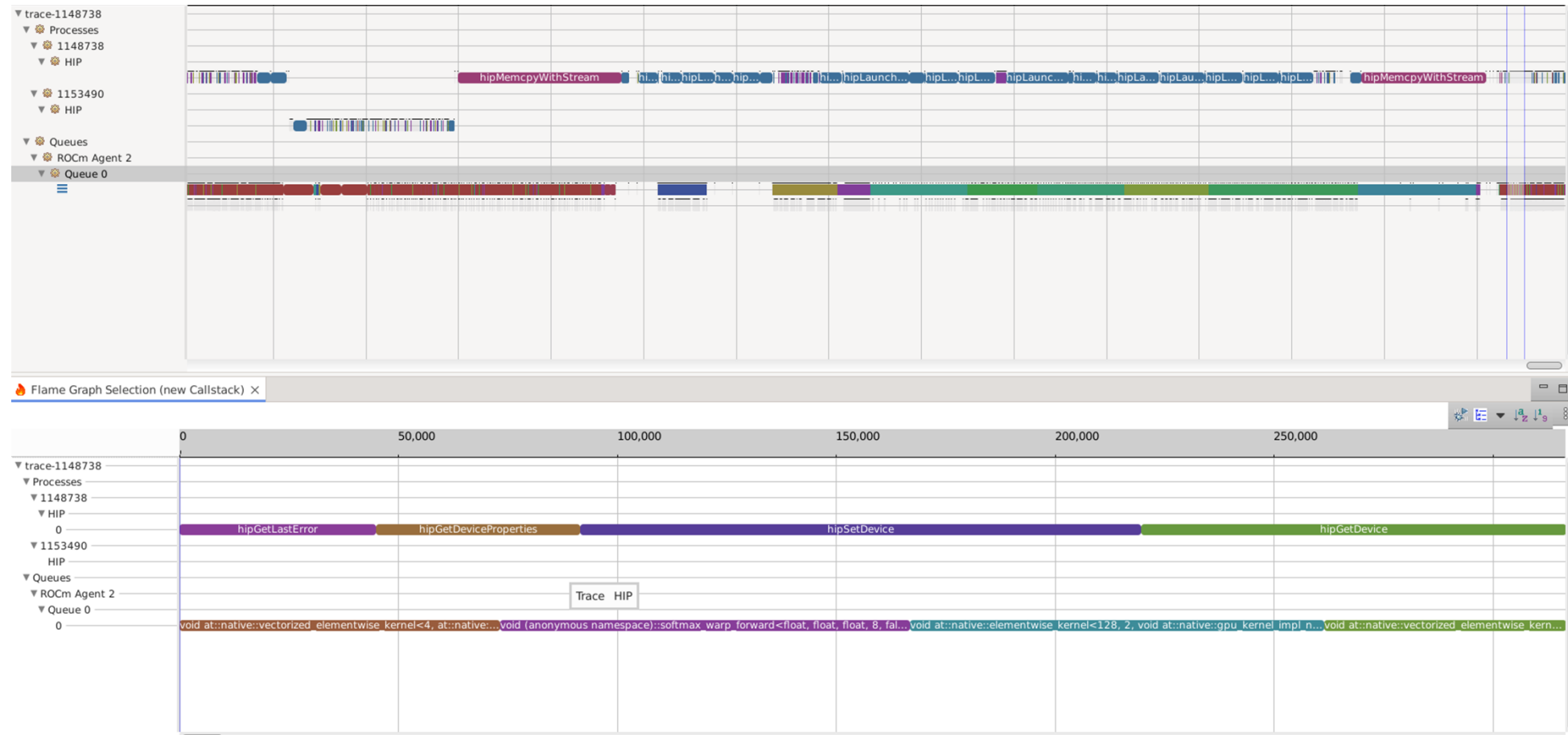
CTF format trace



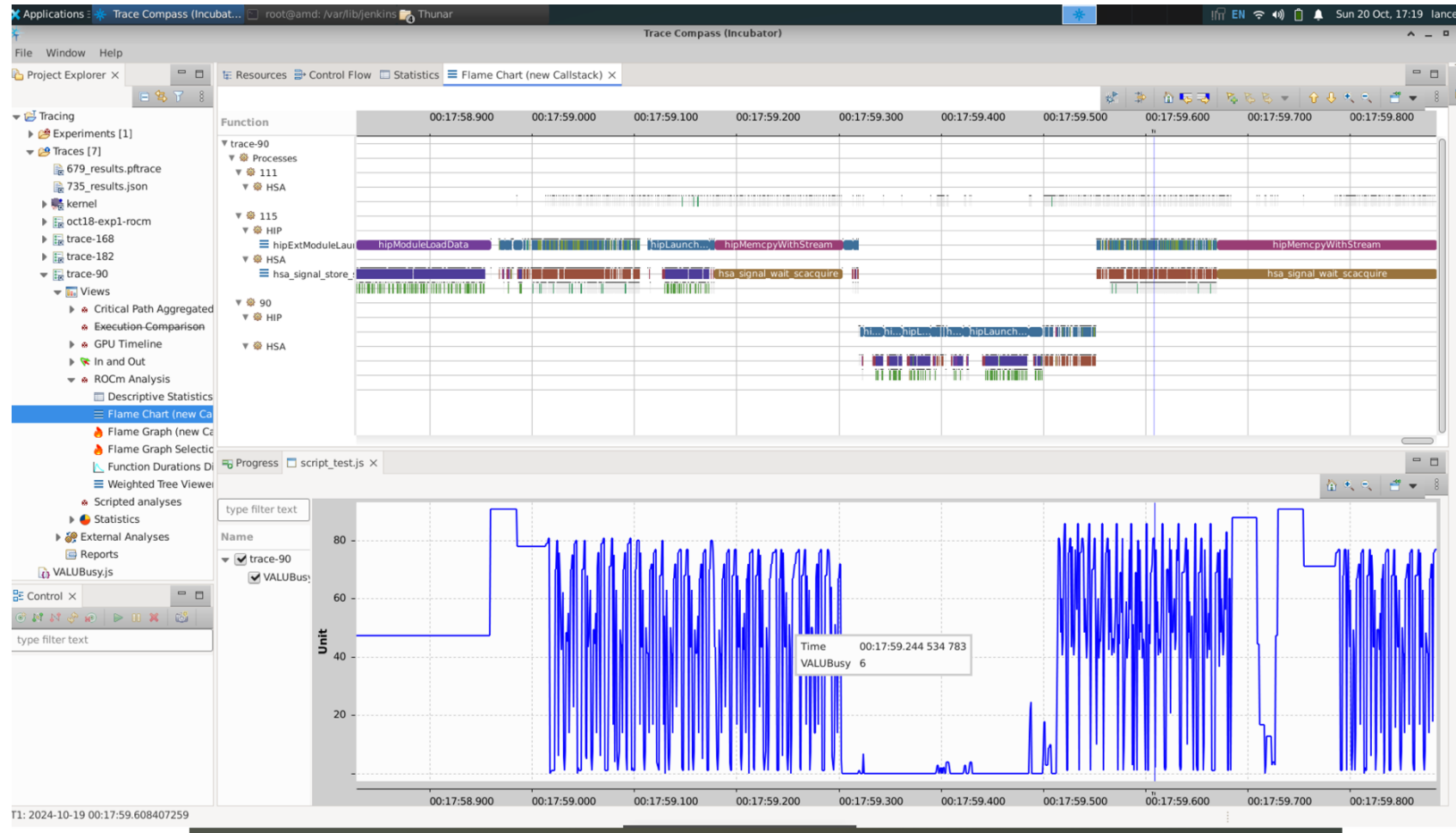
Nvidia trace



Analysis of a single Iteration of Bloom model



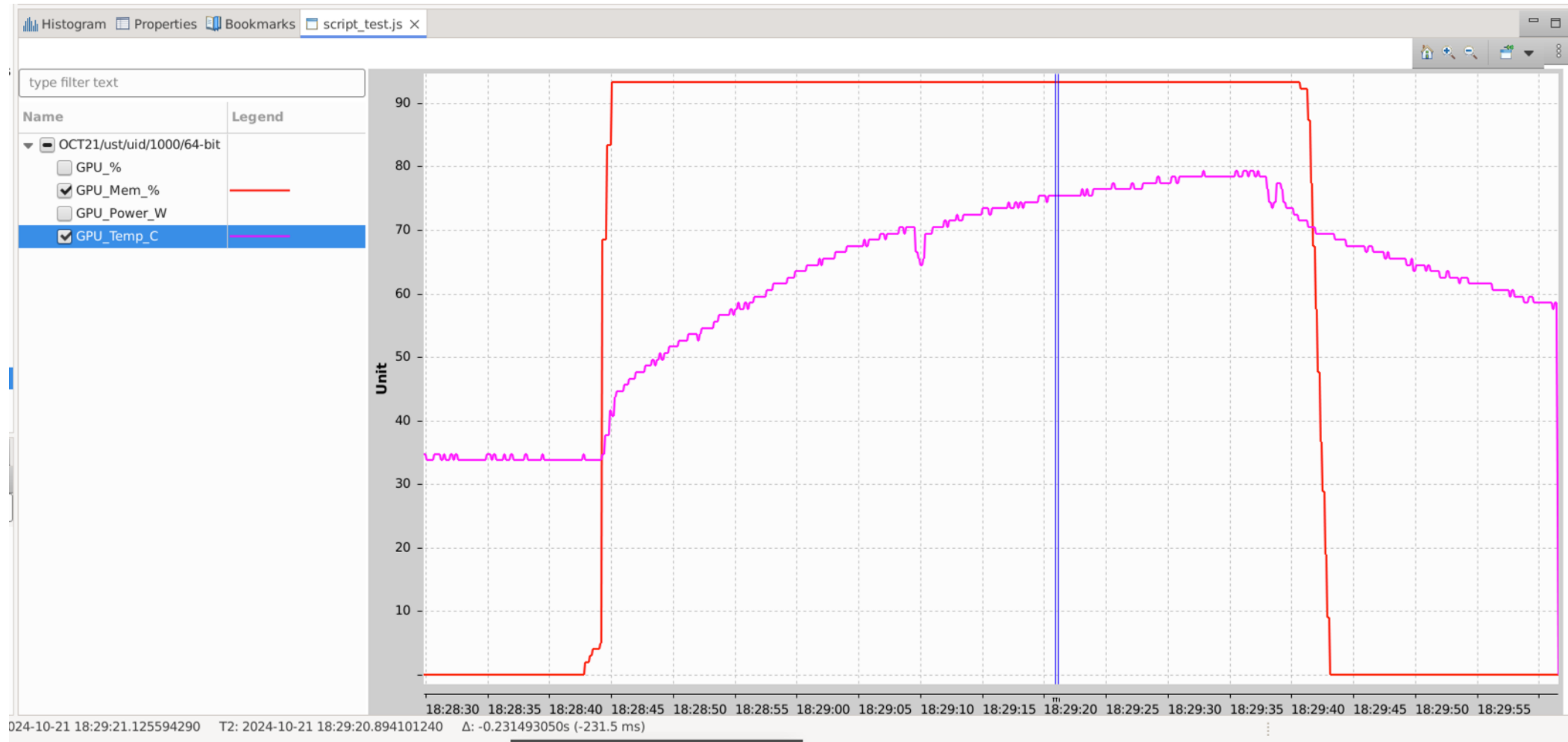
Integration of GPU counter data in TC with scripts



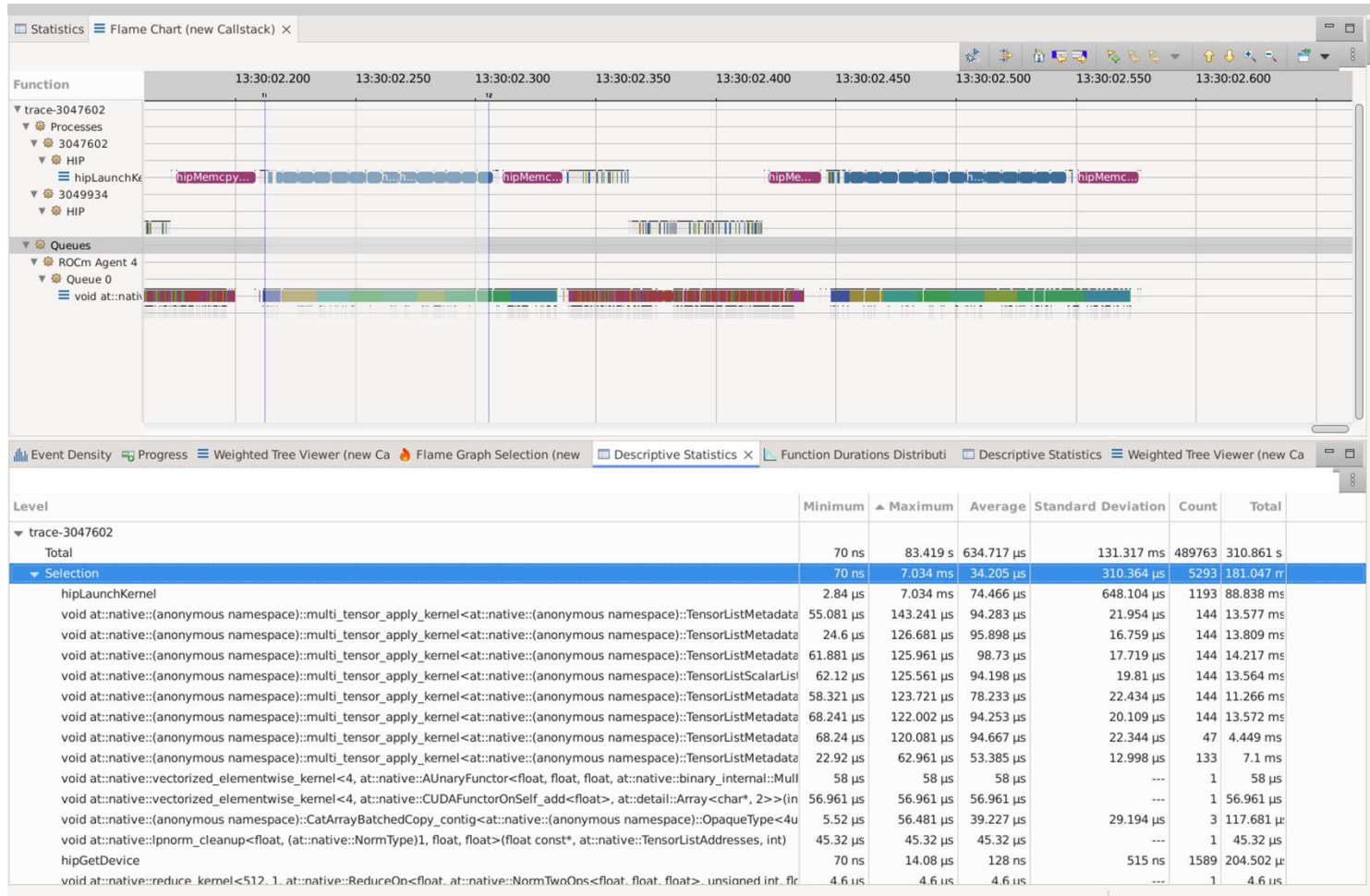
Requires LLTNG ust
Requires extra script

Not suitable for non-root env
common in HPC

Scripting with TC to add GPU metrics



Visualizing Hip Activity of Queue 0



• **TODO:**

Add a Pie Chart

Display the functions without all the signatures

Future works

- Human Readable names for functions in Queue 0
- Integrated stats view on for events in Queue 0
- Comparative analysis of duration of function calls across different accelerators