



TEXAS ADVANCED COMPUTING CENTER

WWW.TACC.UTEXAS.EDU



TEXAS

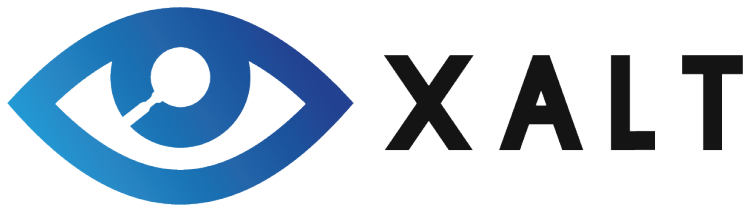
The University of Texas at Austin

How Sampling changed XALT and why Signals won't save us

Robert McLay

July 20, 2023

XALT: Outline



- ▶ XALT 2 can collect every execution
- ▶ This is too much data (at least for TACC)
- ▶ XALT 2 supports sampling to reduce the firehose of data.
- ▶ This brought many changes in the way XALT works
- ▶ I had hoped that signalling would allow XALT to drop the start record.

Too much data

- ▶ One job used two nodes to generate 2 Million records
- ▶ It took over 4 days to load the 2 million records
- ▶ Then another job used small 2 node executions to train a neural network.
- ▶ We were drowning in data

Too much data \Rightarrow Sampling

- ▶ Old XALT generated a start and end record for all executions
- ▶ This way failed executions could be tracked
- ▶ Site controlled sampling based on runtime.
- ▶ The longer the execution the more likely it will be “tracked” or recorded.

Start Record?

- ▶ XALT doesn't generate a start record for NON-MPI executions
- ▶ Do not want a start record that has to be ignored
- ▶ Problem: What about long running MPI programs that terminate by Job Scheduler
- ▶ No End Record
- ▶ Want to track these executions

“Large” MPI programs special treatment

- ▶ Any execution $\text{MPI_Tasks} < \text{MPI_ALWAYS_RECORD}$ (128 at TACC) will NOT generate a start record
- ▶ $\text{MPI_Tasks} \geq \text{MPI_ALWAYS_RECORD}$ will generate a start record
- ▶ This way “Big” executions will always be tracked.
- ▶ If runtime = zero then use job endtime to record runtime.
- ▶ This is an extra steps that sites must do
- ▶ This is outside of XALT data.

Upshot of these rules

- ▶ XALT knows nothing about non-MPI executions that fail
- ▶ XALT knows nothing about “small” MPI executions that fail
- ▶ XALT knows nothing about non-tracked execution

What about signals?

- ▶ Use signals and get rid of the start record for “Large” MPI executions?
- ▶ SLURM sends a signal that a job is about to end.
- ▶ Can we use that?

What about signals? (II)

- ▶ Well, none of the MPI libraries passed that signal through to the running program.
- ▶ Even if it did, XALT would require the signal that was raised on task zero.

Conclusions

- ▶ Sampling helps deal with the firehose of data
- ▶ But Signals won't save us from having to generate a start record on "Large" MPI executions.

Future Topics?

- ▶ Next Meeting will be on August 17, 2023 at 10:00 am U.S. Central (15:00 UTC)