FERMILAB-SLIDES-24-0268-CSAID

Geant4 Computing Performance Task : Protocol Evolution

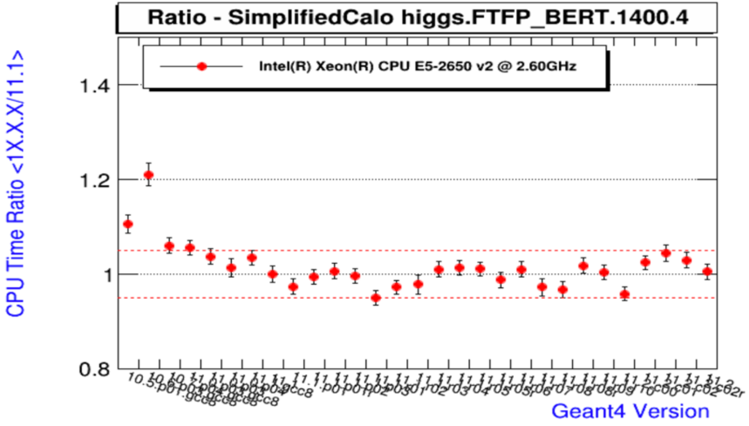
Julia Yarba and Soon Yung Jun (Fermilab)

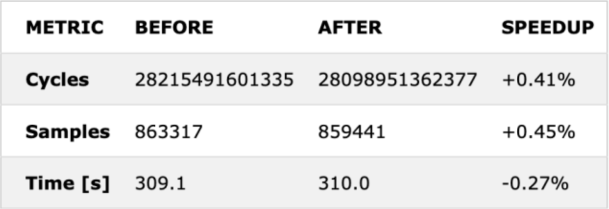
The 29th Geant4 Collaboration Meeting, Catania

Oct. 7 - 11, 2024

Geant4 Computing Performance Task (G4CPT)

- Purpose
  - Monitor CPU & memory through the development cycle
  - Identify issues (if any) and
  - Identify opportunities for code improvement
  - Provide feedback to the working group leaders
  - Close all open issues before the next release
- Ongoing activities
  - Regular profiling/benchmarking of Geant4 development and public release, specific development tags as needed (total 20+ rounds per year, 50+ test samples per round, each sample runs multiple times to define mean and error)
  - Performance difference report from CI, triggered by the merge request (1 test app runs once per monitoring round)



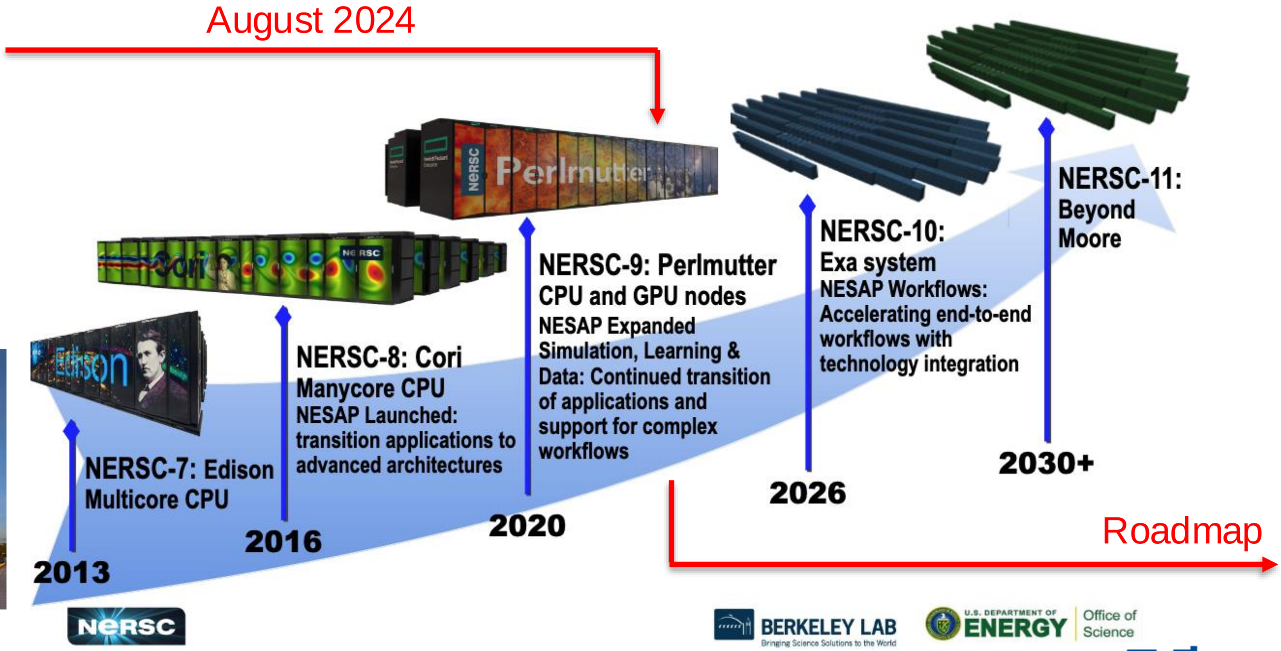


METRIC	BEFORE	AFTER	SPEEDUP
Cycles	28215491601335	28098951362377	+0.41%
Samples	863317	859441	+0.45%
Time [s]	309.1	310.0	-0.27%

Resources: Migration(s) 2024

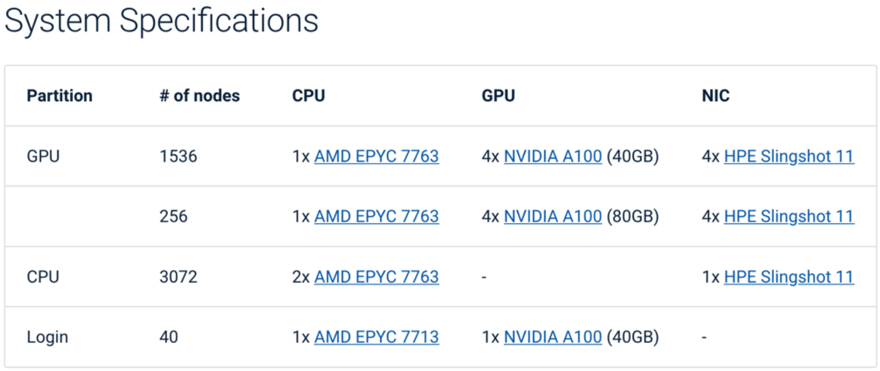
- Wilson Cluster at Fermilab (IntelXeonCPUE52650@2.60GHz) – from SL7 to EL8, Jan. 2024
- NERSC at LBNL (Perlmutter, AMD EPYC 7713 64-Core Processor, SUSE Linux)

Wilson Cluster @Fermilab


August 2024
NERSC-7: Edison
Multicore CPU
2013
NERSC-8: Cori
Manycore CPU
NESAP Launched: transition applications to advanced architectures
2016
NERSC-9: Perlmutter
CPU and GPU nodes
NESAP Expanded Simulation, Learning & Data: Continued transition of applications and support for complex workflows
2020
NERSC-10:
Exa system
NESAP Workflows: Accelerating end-to-end workflows with technology integration
2026
NERSC-11:
Beyond Moore
2030+
Roadmap
NERSC
BERKELEY LAB
U.S. DEPARTMENT OF ENERGY
Office of Science

Overview of NERSC Resources

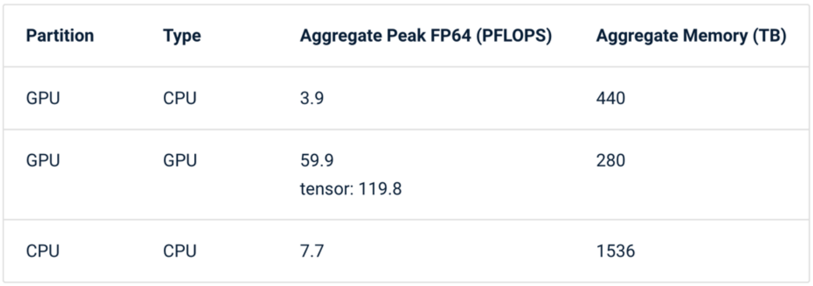
- NERSC is **N**ational **E**nergy **R**esearch **S**cientific **C**omputing Center
  - High Performance Computing and Storage facilities and support for research sponsored by, and of interest to, the U.S. Department of Energy Office of Science
- Perlmutter: HPE (Hewlett Packard Enterprise) Cray EX supercomputer
- Based on the HPE Cray Shasta platform
- Hybrid system of 3,072 CPU-only and 1,792 GPU-accelerated nodes



System Specifications

Partition	# of nodes	CPU	GPU	NIC
GPU	1536	1x AMD EPYC 7763	4x NVIDIA A100 (40GB)	4x HPE Slingshot 11
	256	1x AMD EPYC 7763	4x NVIDIA A100 (80GB)	4x HPE Slingshot 11
CPU	3072	2x AMD EPYC 7763	-	1x HPE Slingshot 11
Login	40	1x AMD EPYC 7713	1x NVIDIA A100 (40GB)	-

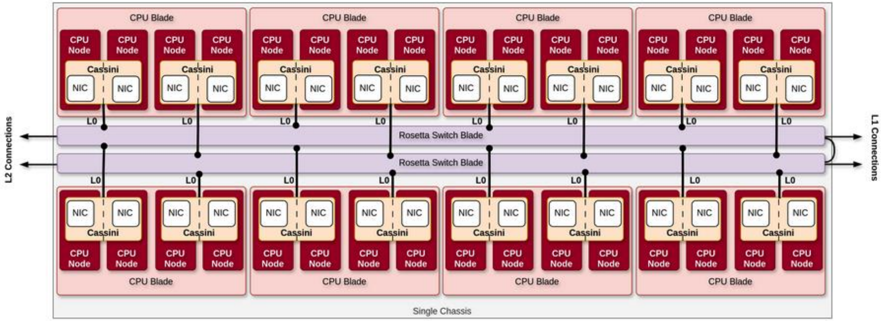
System Performance (79 PFlop/s; Rank 14, Jun 2024)

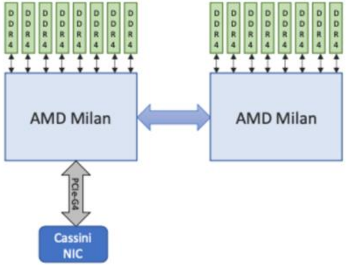


Partition	Type	Aggregate Peak FP64 (PFLOPS)	Aggregate Memory (TB)
GPU	CPU	3.9	440
GPU	GPU	59.9 tensor: 119.8	280
CPU	CPU	7.7	1536

Perlmutter CPU nodes

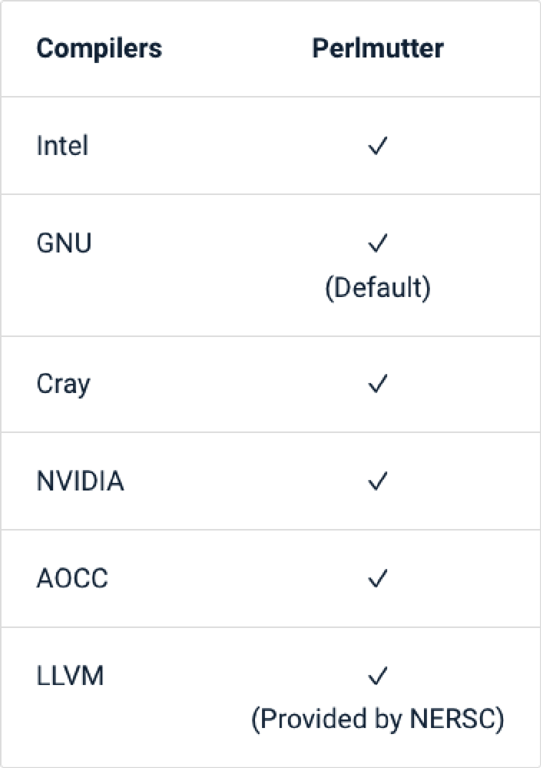
- Specification of CPU nodes
  - 2x AMD EPYC 7763 (Milan) CPUs
  - 64 cores per CPU (2 threads/core, 256 total)
  - AVX2 instruction set
  - 512 GB of DDR4 memory total
  - 204.8 GB/s memory bandwidth per CPU
  - 1x HPE Slingshot 11 NIC
  - PCIe 4.0 NIC-CPU connection
  - 39.2 GFlops per core
  - 2.51 TFlops per socket
  - 4 NUMA domains per socket (NPS=4)





Compilers and Build Tools

- Compilers used recently
  - gcc11-series up to July 2024 (Wilson, SL7 & EL8)
  - gcc12.3.0 from August 2024 (NERSC, SUSE Linux)
- Available compilers on Perlmutter
  - Basic compilers
  - Compiler wrappers
- Build tools
  - cmake
  - spack
- Resource management and job scheduling
  - SLURM

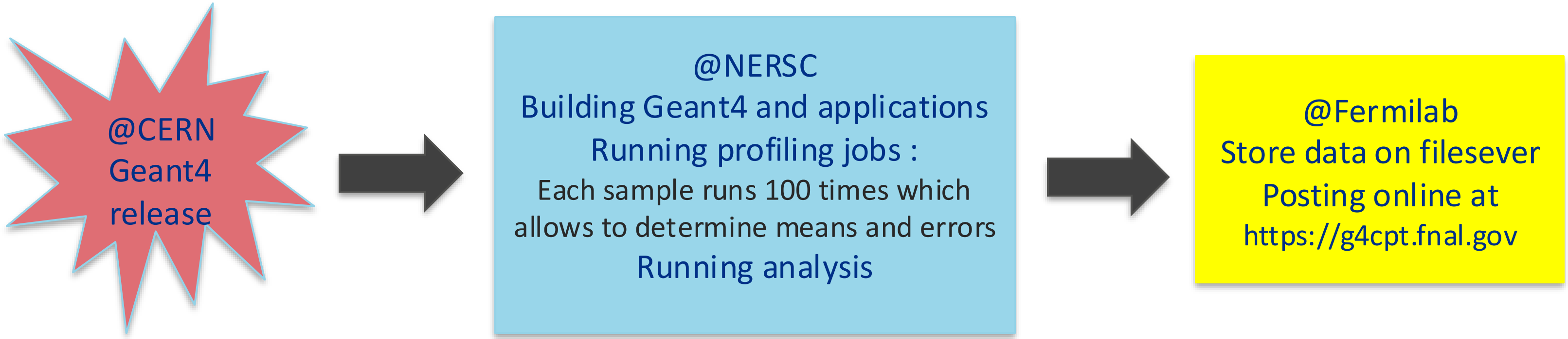


Compilers	Perlmutter
Intel	✓
GNU	✓ (Default)
Cray	✓
NVIDIA	✓
AOCC	✓
LLVM	✓ (Provided by NERSC)

Profiling Tools

- Current Profiling tools
  - Open|SpeedShop (https://github.com/OpenSpeedShop)
    - 2.4.1 → development revision (for compatibility with modern platforms)
    - No free support any longer (→ licensed SurveyPerf by Trenza Synergy; although still open source)
  - IgProf 5.9.18 : incompatible with EL8 but reinstated at NERSC, multithreading is not supported
- NERSC provides other popular profiling tools
  - Codee (previously known as Parallelware Analyzer)
  - CrayPat (performance analysis tool offered by Cray)
  - Darshan (extended tracing module, DXT)
  - MAP, part of the Linaro Forge (previously known as Arm Forge or Allinea Forge) tool suite
  - NVIDIA® Nsight™ Systems
- Other tools under consideration: HPCToolkits, etc.

Overview of Profiling Round



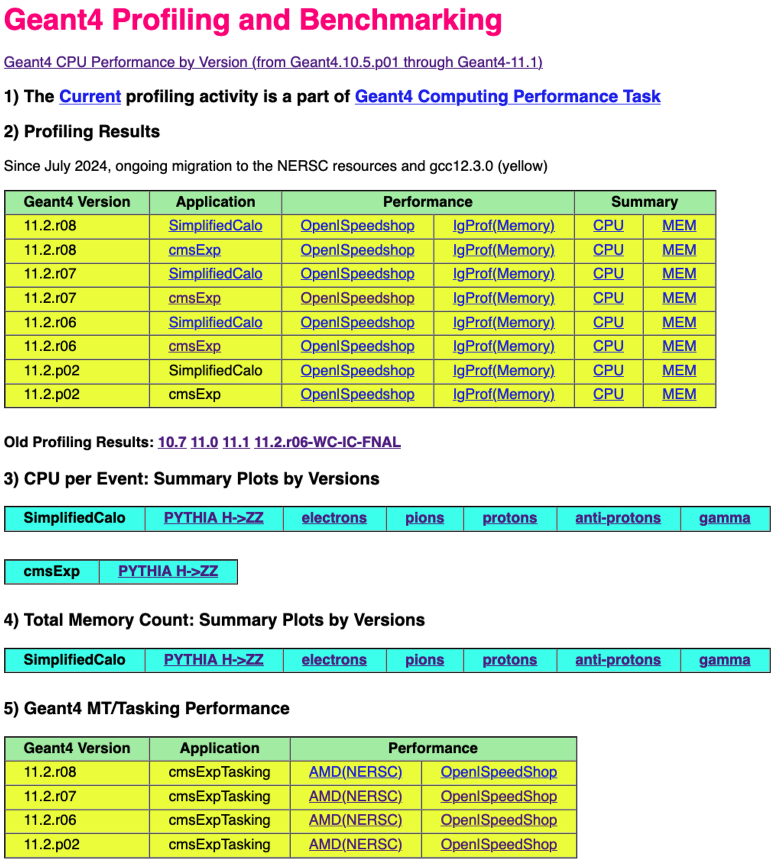
@CERN
Geant4
release

@NERSC
Building Geant4 and applications
Running profiling jobs :
Each sample runs 100 times which
allows to determine means and errors
Running analysis

@Fermilab
Store data on filesever
Posting online at
https://g4cpt.fnal.gov

- Running batch jobs at NERSC
  - Computing resources under the m4599 project (Fermilab IF at NERSC)
  - Currently G4CPT has the annual CPU quota of 3500 node-hours (renewable) - ~25 rounds
  - Reserved nodes (a week in advance)
  - Premium queues on-demand (double the CPU cost)

Delivering Results (https://g4cpt.fnal.gov)



Geant4 Profiling and Benchmarking

Geant4 CPU Performance by Version (from Geant4.10.5.p01 through Geant4-11.1)

1) The Current profiling activity is a part of Geant4 Computing Performance Task

2) Profiling Results

Since July 2024, ongoing migration to the NERSC resources and gcc12.3.0 (yellow)

Geant4 Version	Application	Performance		Summary	
11.2.r08	SimplifiedCalo	OpenISpeedshop	IgProf(Memory)	CPU	MEM
11.2.r08	cmsExp	OpenISpeedshop	IgProf(Memory)	CPU	MEM
11.2.r07	SimplifiedCalo	OpenISpeedshop	IgProf(Memory)	CPU	MEM
11.2.r07	cmsExp	OpenISpeedshop	IgProf(Memory)	CPU	MEM
11.2.r06	SimplifiedCalo	OpenISpeedshop	IgProf(Memory)	CPU	MEM
11.2.r06	cmsExp	OpenISpeedshop	IgProf(Memory)	CPU	MEM
11.2.p02	SimplifiedCalo	OpenISpeedshop	IgProf(Memory)	CPU	MEM
11.2.p02	cmsExp	OpenISpeedshop	IgProf(Memory)	CPU	MEM

Old Profiling Results: 10.7 11.0 11.1 11.2.r06-WC-IC-FNAL

3) CPU per Event: Summary Plots by Versions

SimplifiedCalo	PYTHIA H->ZZ	electrons	pions	protons	anti-protons	gamma

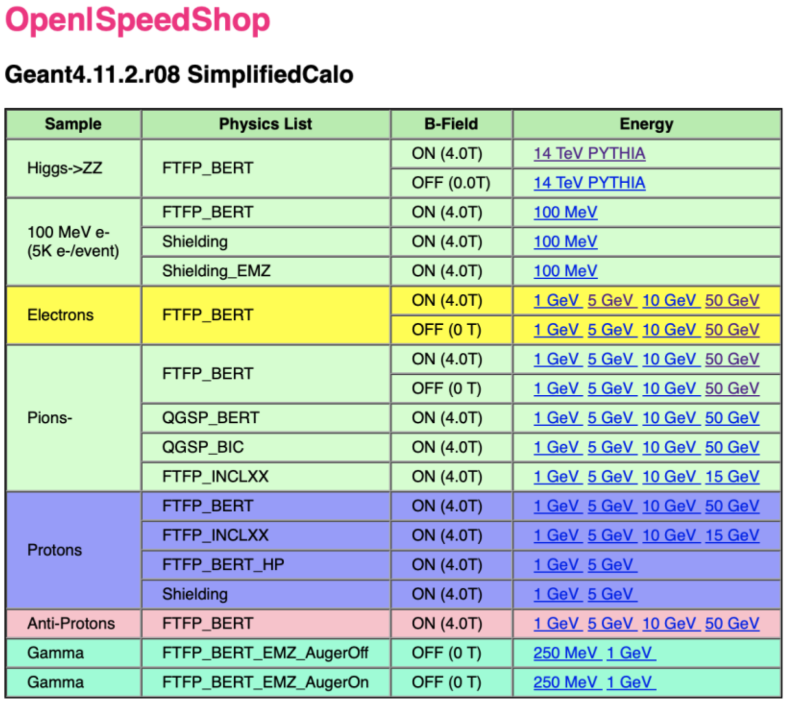
cmsExp	PYTHIA H->ZZ

4) Total Memory Count: Summary Plots by Versions

SimplifiedCalo	PYTHIA H->ZZ	electrons	pions	protons	anti-protons	gamma

5) Geant4 MT/Tasking Performance

Geant4 Version	Application	Performance	
11.2.r08	cmsExpTasking	AMD(NERSC)	OpenISpeedShop
11.2.r07	cmsExpTasking	AMD(NERSC)	OpenISpeedShop
11.2.r06	cmsExpTasking	AMD(NERSC)	OpenISpeedShop
11.2.p02	cmsExpTasking	AMD(NERSC)	OpenISpeedShop



OpenISpeedShop

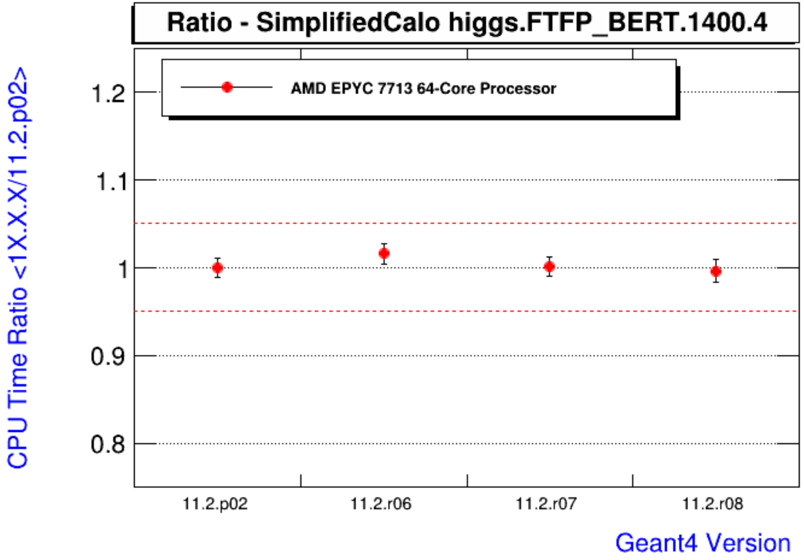
Geant4.11.2.r08 SimplifiedCalo

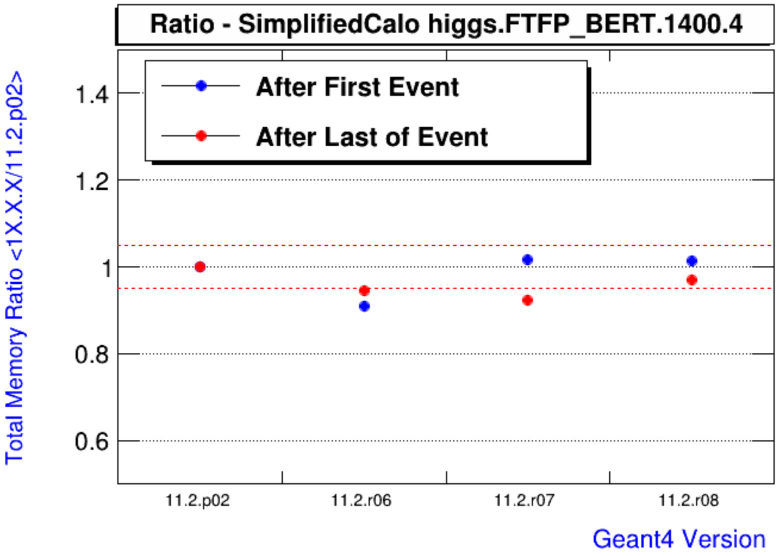
Sample	Physics List	B-Field	Energy
Higgs->ZZ	FTFP_BERT	ON (4.0T)	14 TeV PYTHIA
		OFF (0.0T)	14 TeV PYTHIA
100 MeV e- (5K e-/event)	FTFP_BERT	ON (4.0T)	100 MeV
	Shielding	ON (4.0T)	100 MeV
	Shielding_EMZ	ON (4.0T)	100 MeV
Electrons	FTFP_BERT	ON (4.0T)	1 GeV 5 GeV 10 GeV 50 GeV
		OFF (0 T)	1 GeV 5 GeV 10 GeV 50 GeV
Pions-	FTFP_BERT	ON (4.0T)	1 GeV 5 GeV 10 GeV 50 GeV
		OFF (0 T)	1 GeV 5 GeV 10 GeV 50 GeV
	QGSP_BERT	ON (4.0T)	1 GeV 5 GeV 10 GeV 50 GeV
	QGSP_BIC	ON (4.0T)	1 GeV 5 GeV 10 GeV 50 GeV
	FTFP_INCLXX	ON (4.0T)	1 GeV 5 GeV 10 GeV 15 GeV
Protons	FTFP_BERT	ON (4.0T)	1 GeV 5 GeV 10 GeV 50 GeV
	FTFP_INCLXX	ON (4.0T)	1 GeV 5 GeV 10 GeV 15 GeV
	FTFP_BERT_HP	ON (4.0T)	1 GeV 5 GeV
	Shielding	ON (4.0T)	1 GeV 5 GeV
Anti-Protons	FTFP_BERT	ON (4.0T)	1 GeV 5 GeV 10 GeV 50 GeV
Gamma	FTFP_BERT_EMZ_AugerOff	OFF (0 T)	250 MeV 1 GeV
Gamma	FTFP_BERT_EMZ_AugerOn	OFF (0 T)	250 MeV 1 GeV

We believe that, in general, we reasonably cover all aspects that can be critical for the Geant4 development. However, feedback from representatives from experiments and projects is welcome, in case we miss something important.

CPU and Memory Trends in Geant4

- Recent benchmarking on Perlmutter (after the August 2024 migration)





The number of steps and tracks (geometry vs physics) are also measured, e.g.

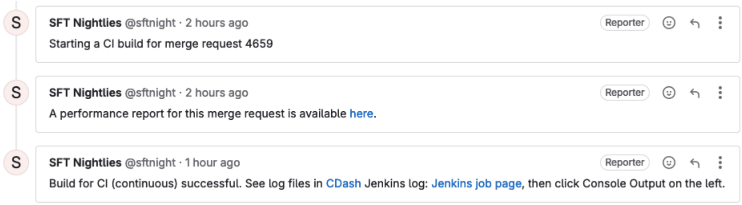
https://g4cpt.fnal.gov/g4p/oss_11.2.r08_SimplifiedCalo_01/higgs.FTFP_BERT.1400.4/prof_nstep_particle_list.html

Observed Issues to Resolve

- Identified issues to improve measurements
  - Larger than desired CPU measurement errors but gradually improving
    - Examples in backup
    - Core-to-core fluctuation ? Pinning ? NUMA effect ?
  - Occasional out-of-memory job failure
    - Random, towards the end of jobs; I/O interruption ?
  - Find the optimal number of cores per sample (controlling jobs efficiently vs. optimal use of resources)
- Optimal use of allocated CPU quota
  - Allocated quota is **node-hours** based (total 128 cores, or 256 threads, aka logical CPUs, per node)
  - Optimal run schedules for full occupancy of the reserved resources, to avoid idle time

Performance Monitor Report by Merge Request (MR)

- Automatic performance report integrated into Jenkins (work by G. Amadio)
  - Runs for each merge request opened for Geant4 (since 2022)

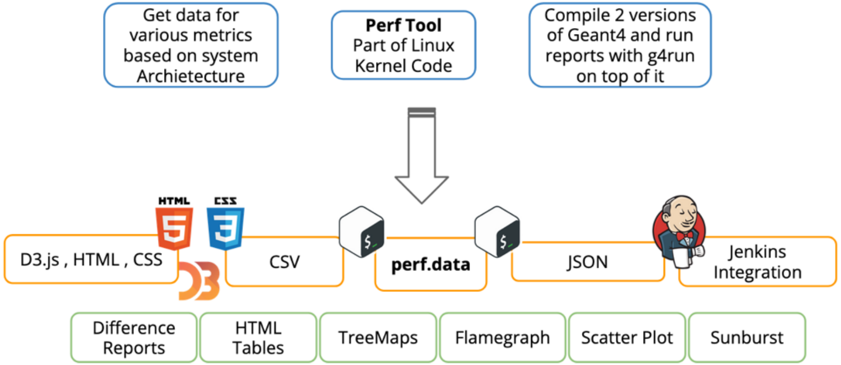


SFT Nightlies @sftnight · 2 hours ago
Starting a CI build for merge request 4659

SFT Nightlies @sftnight · 2 hours ago
A performance report for this merge request is available here.

SFT Nightlies @sftnight · 1 hour ago
Build for CI (continuous) successful. See log files in CDash Jenkins log: Jenkins job page, then click Console Output on the left.

  - Identify performance regressions as they happen



Get data for various metrics based on system Archietecture

Perf Tool Part of Linux Kernel Code

Compile 2 versions of Geant4 and run reports with g4run on top of it

D3.js , HTML , CSS

CSV

perf.data

JSON

Jenkins Integration

Difference Reports

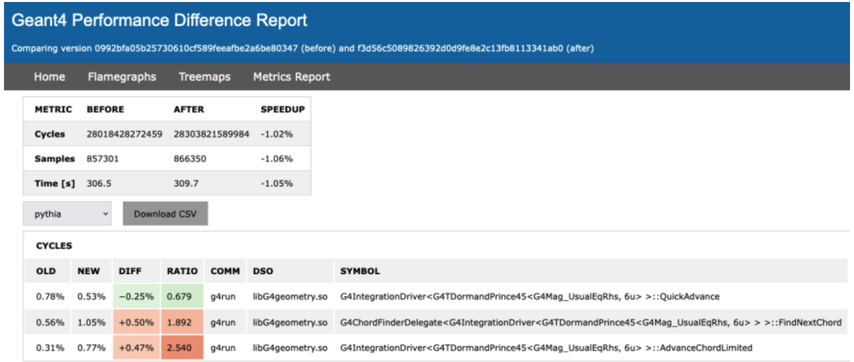
HTML Tables

TreeMaps

Flamegraph

Scatter Plot

Sunburst



Geant4 Performance Difference Report

Comparing version 0992bfa05b25730610cf589feeafbe2a6be80347 (before) and f3d56c5089826392d0d9fe8e2c13fb8113341ab0 (after)

METRIC	BEFORE	AFTER	SPEEDUP
Cycles	28018428272459	28303821589984	-1.02%
Samples	857301	866350	-1.06%
Time [s]	306.5	309.7	-1.05%

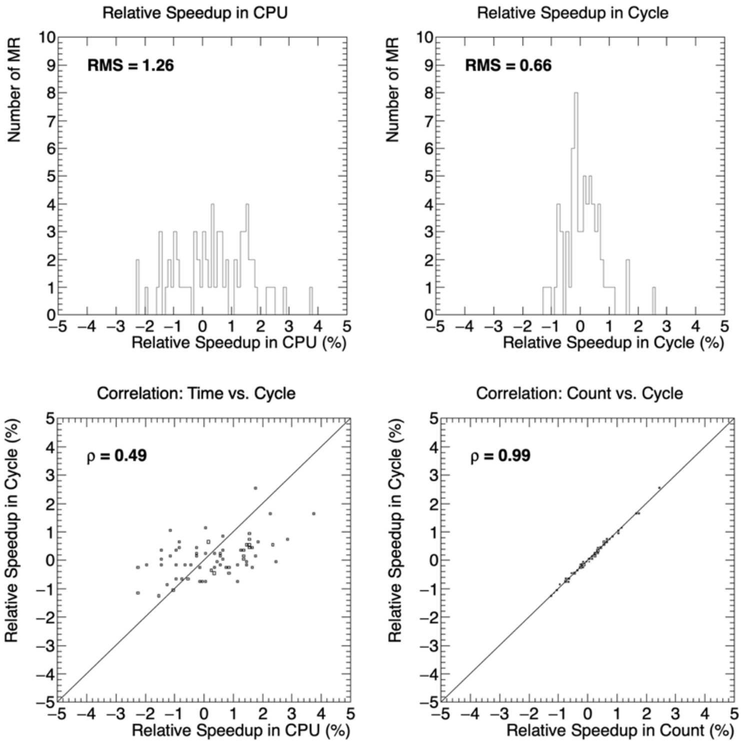
CYCLES

OLD	NEW	DIFF	RATIO	COMM	DSO	SYMBOL
0.78%	0.53%	−0.25%	0.679	g4run	libG4geometry.so	G4IntegrationDriver<G4TDormandPrince45<G4Mag_UsualEqRhs, 6u> >::QuickAdvance
0.56%	1.05%	+0.50%	1.892	g4run	libG4geometry.so	G4ChordFinderDelegate<G4IntegrationDriver<G4TDormandPrince45<G4Mag_UsualEqRhs, 6u> > >::FindNextChord
0.31%	0.77%	+0.47%	2.540	g4run	libG4geometry.so	G4IntegrationDriver<G4TDormandPrince45<G4Mag_UsualEqRhs, 6u> >::AdvanceChordLimited

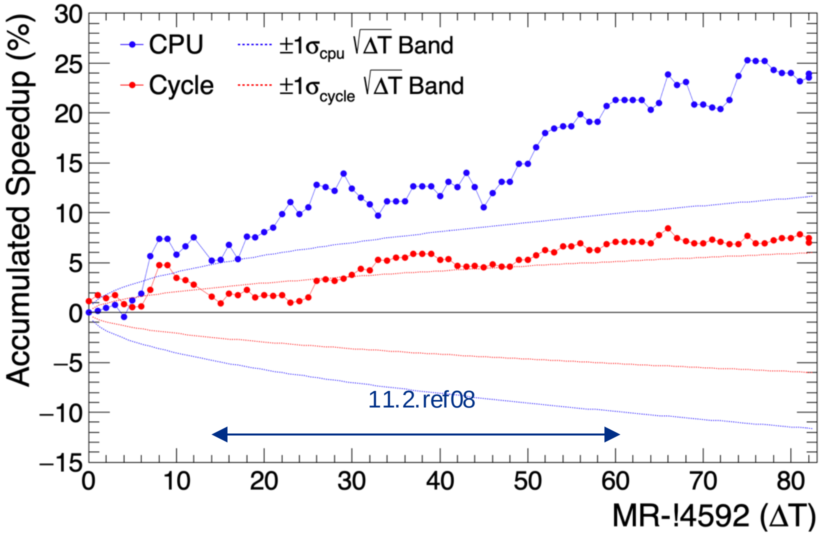
NOTE: Speedup is performance difference comparing the master (before) and the MR branch (after).

Regression of Performance Monitor Report

- Preserved data since Aug. 22, 2024 (MR !4592)



The accumulated gain assuming the fluctuation follows the geometric Brownian diffusion process



Q: Would this analysis be consistent with statistical measurements from reference releases ?

Proposed Plan

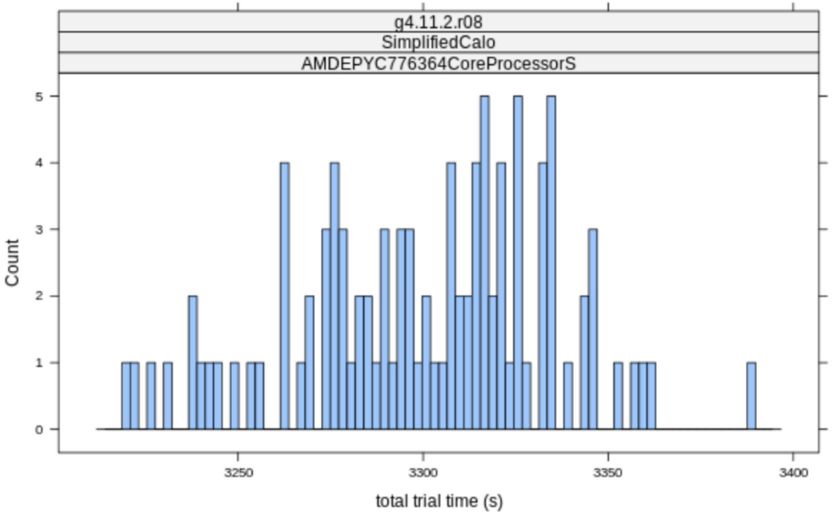
- Fully benefit from both components (CI and monthly) of the performance monitoring system
  - Including estimate of the measurement errors from the CI/MR monitoring
- Continue the initiative, preserve performance data from CI (EOS and/or disk at Fermilab) assuming the post regression analysis may help or serve as complimentary for the monthly measurement
  - The lifetime of output on CI is a week
- Monitor and identify MRs which significantly contributes to performance changes in a particular development cycle
- Communicate with developers for post justifications

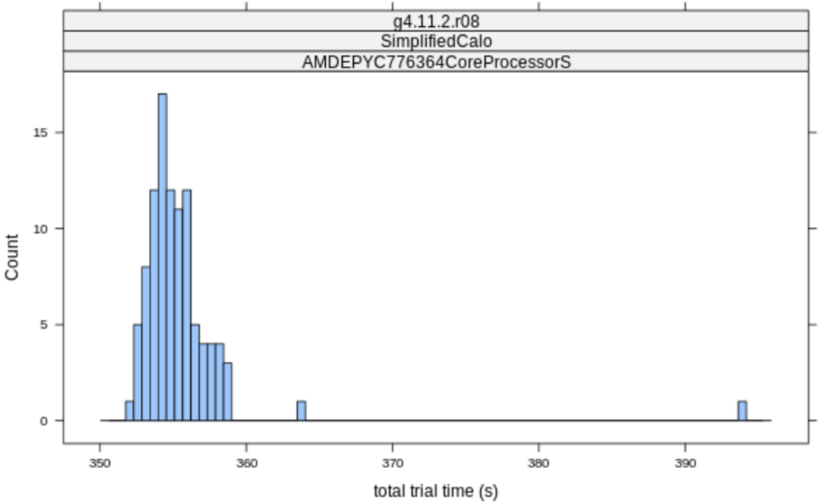
Summary

- Performed CPU and memory profiling for development and public releases
  - Measurements are done on substantial statistical basis
- Reported results to the working group leaders
- Presented results and issues of computing performance at the Steering Board meetings
- Initial regression analysis of the performance monitor report by the merge request
- Work carried out along with two major migrations
  - From SL7 to SL8 on Wilson (now decommissioned) (Jan 2024)
  - From Wilson/Fermilab to Perlmutter/NERSC/LBNL (Aug 2024)
  - **Many thanks to the NERSC team for their support !**

BACKUP SLIDES

Measurements are done on statistical basis



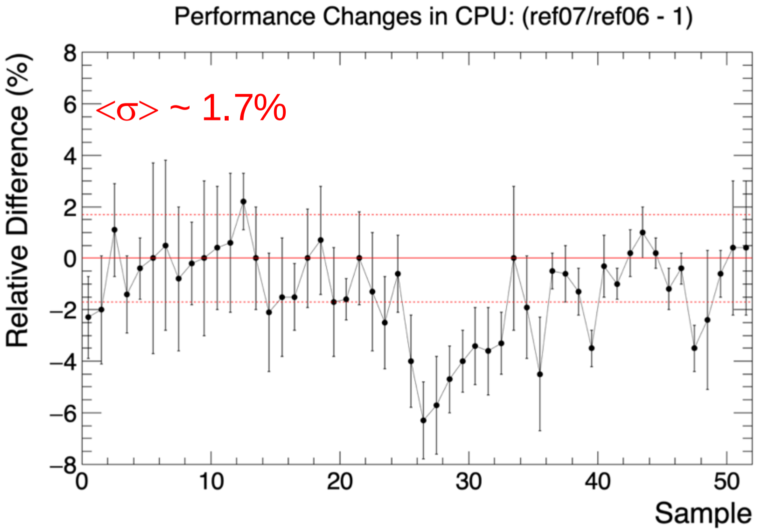


CPU estimate e.g. for Higgs input sample processed through SimplifiedCalo geometry is repeated multiple times which allows to determine mean and error

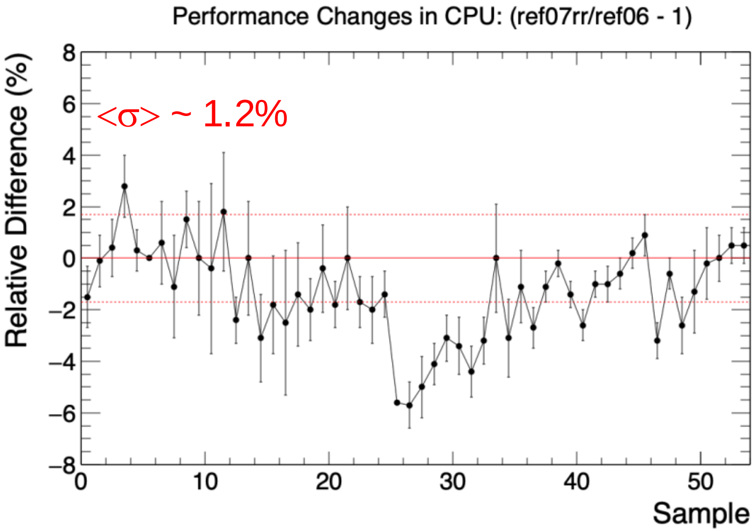
CPU estimate e.g. for single 50 GeV electron on input processed through SimplifiedCalo geometry is repeated multiple times which allows to determine mean and error

First Measurements : geant4.11.2-ref07, errors

- Prepared new references (Geant4 11.2.p02 and 11.2.ref06) (multiple measurements)
- Measurement errors are relatively large but are gradually improving
  - Core by core fluctuations ? (used to be ~1% on Wilson, ~0.8% on tev, 0.5% on cluck)



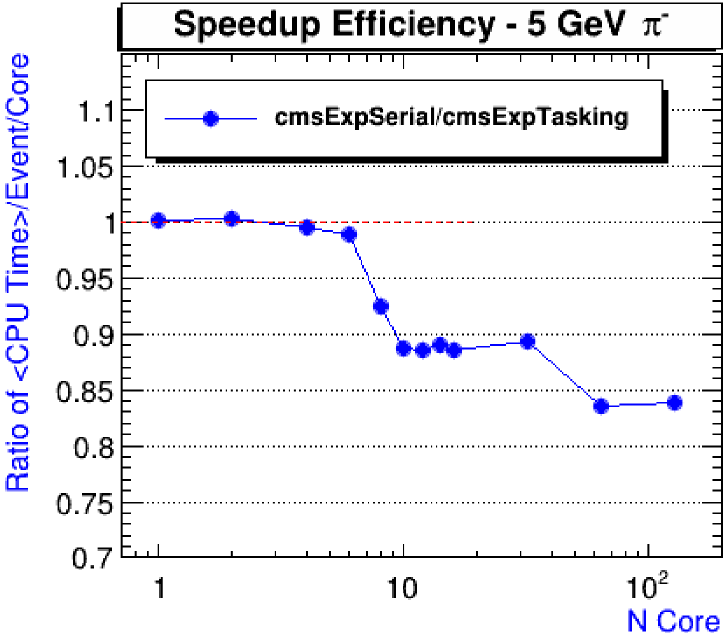
Pinning each process to a core, using 1 thread on a core

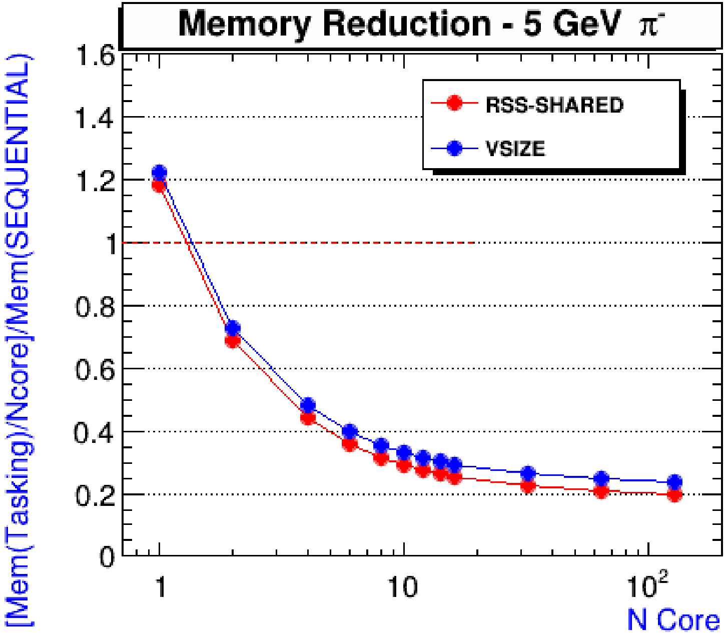


Pinning each process to a to a core, using **both** threads on a core

Geant4-Serial over Geant4-Tasking

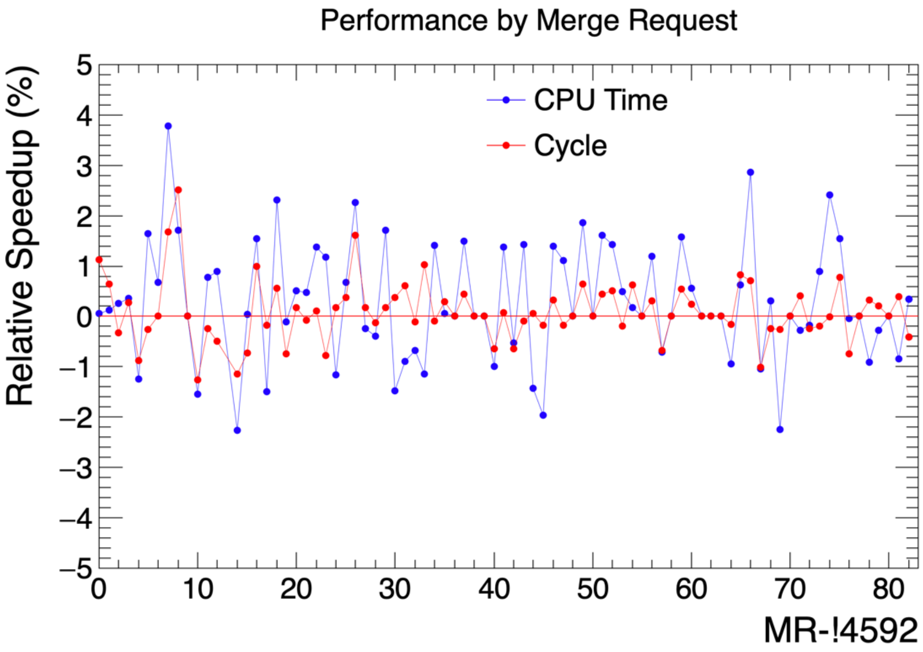
- Multithreading measurements extended to 128 threads; was 16/Intel or 32/AMD (in the past)
  - CPU/Event/core and memory/core as the number of cores

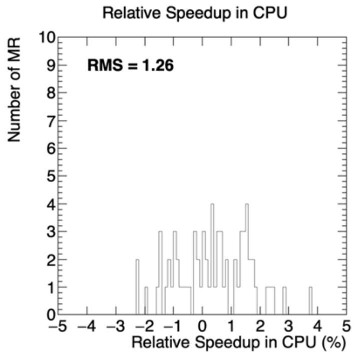


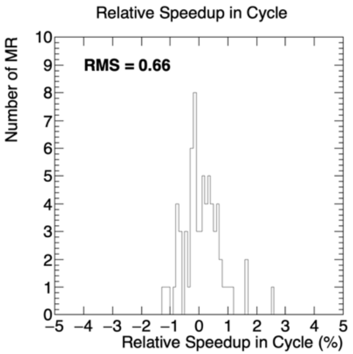


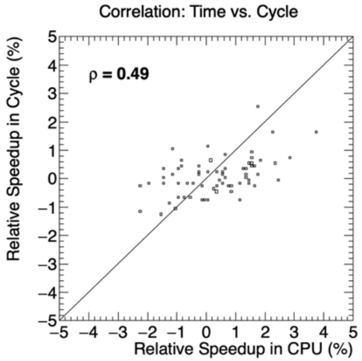
Regression of Performance Monitor Report

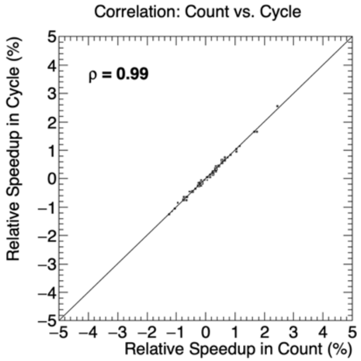
- Records performance data for Geant4 application similar to cmsExp from G4CPT/FNAL
- Data since Aug. 22, 2024 (MR !4592)


Performance by Merge Request
Relative Speedup (%)
MR-!4592
CPU Time
Cycle


Relative Speedup in CPU
Number of MR
RMS = 1.26
Relative Speedup in CPU (%)


Relative Speedup in Cycle
Number of MR
RMS = 0.66
Relative Speedup in Cycle (%)


Correlation: Time vs. Cycle
Relative Speedup in Cycle (%)
ρ = 0.49
Relative Speedup in CPU (%)


Correlation: Count vs. Cycle
Relative Speedup in Cycle (%)
ρ = 0.99
Relative Speedup in Count (%)