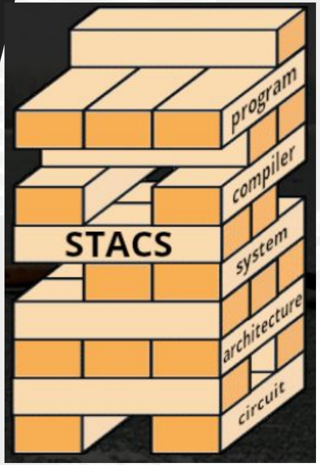
Working from Both Ends to Bridge the Gap Between Silicon and Human Cognition

High Throughput Computing - July 2024

Ranganath (Bujji) Selagamsetty
Robert Klock
Joshua San Miguel
Mikko Lipasti



Outline

- Neuromorphic computing: the bridge between silicon and biology
- Top-down: Drawing inspiration from the auditory cortex for DNS
  - Broad design exploration of network parameters (CHTC GPUs)
- Bottom-up: Improving current CPU architectures for stochastic workloads
  - Characterisation of Random Number Generation schemes (CHTC CPUs)
- Future Work and Opinions

Neuromorphic Computing

- "The opportunity lies in combining the best of biology and silicon"

- Approaches from different directions:
  - Top-down: understand the brain for better algorithms
  - Bottom-up: accelerate existing computing systems for cognitive programs

	Biology	Silicon
Speed	1 msec	1 nsec
Size	1µm - 10µm	10nm - 100nm
Voltage	~ 0.1V	V_{dd} ~1.0V
Neuron Density	100K/mm^2	5k/mm^2
Reliability	80%	< 99.9999%
Synaptic Error Rate	75%	~ 0%
Fan-out (-in)	10^3-10^4	3-4
Dimensions	Pseudo 3D	Pseudo 3D
Synaptic Op Energy	~ 2 fJ	~10pJ
Total Energy	10 Watt	>>10^3 Watt
Temperature	36C - 38C	5C - 60C
Noise effect	Stochastic Resonance	Bad
Criticality	Edge	Far

* Table 1 from Schuller, Ivan K., Stevens, Rick, Pino, Robinson, and Pechan, Michael. Neuromorphic Computing – From Materials Research to Systems Architecture Roundtable. United States: N. p., 2015. Web. doi:10.2172/1283147.

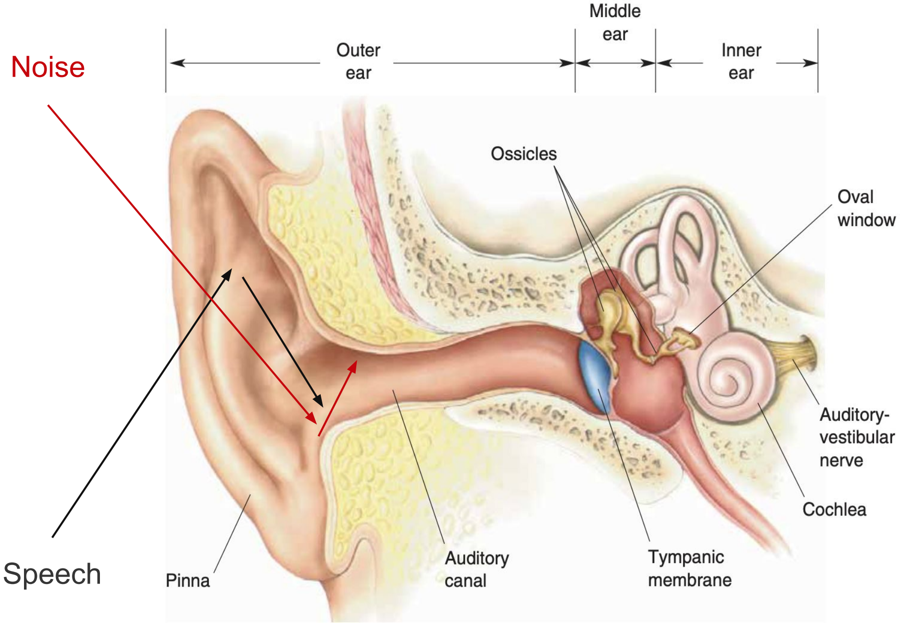
Top-down: Study Auditory Cortex for DNS

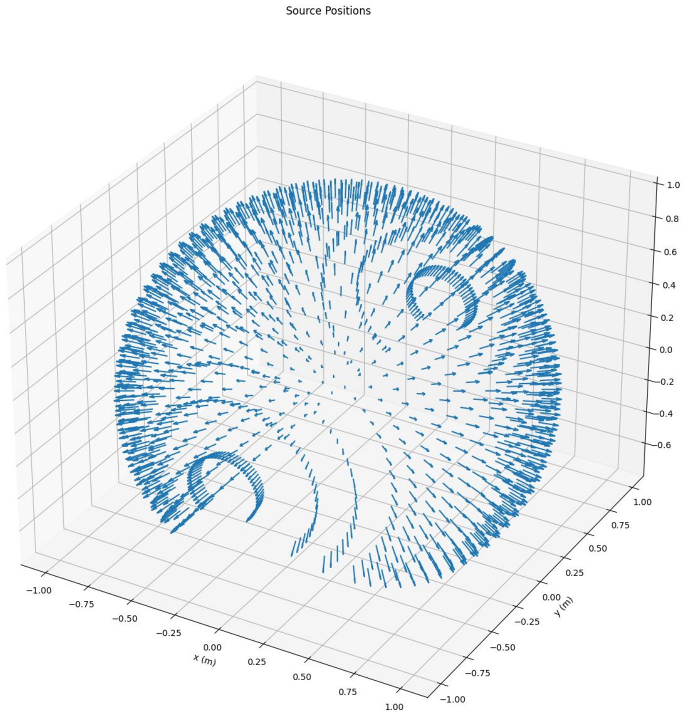
- Speech denoising is a non-trivial, popular task
  - Microsoft DNS
  - Intel N-DNS

- ANNs struggle, ears are proficient
  - Look to human anatomy for inspiration

- What inspiration can we glean from the brain?
  - Rich data encoding from the pinna…
  - Energy efficiency from temporal computing in spiking neural networks…

Speech & Noise Position for Denoising

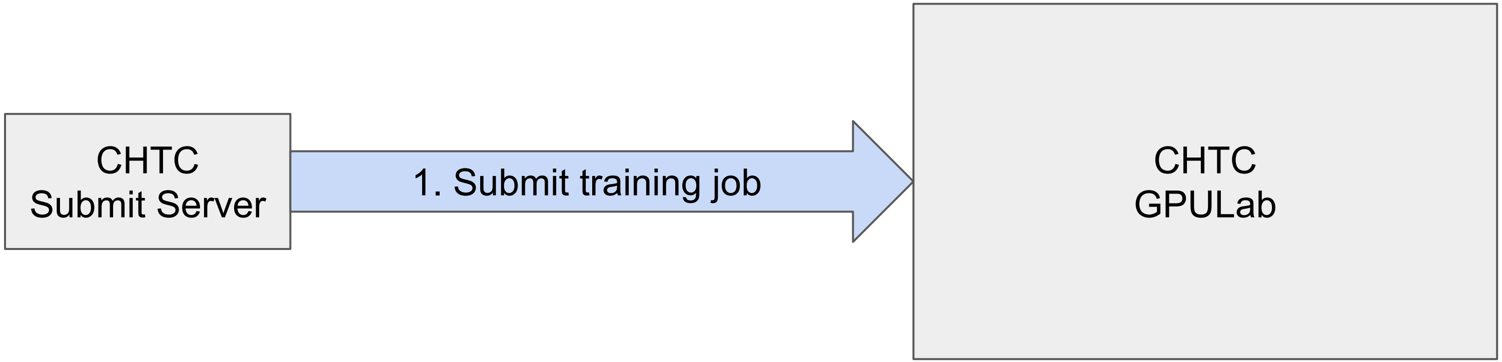




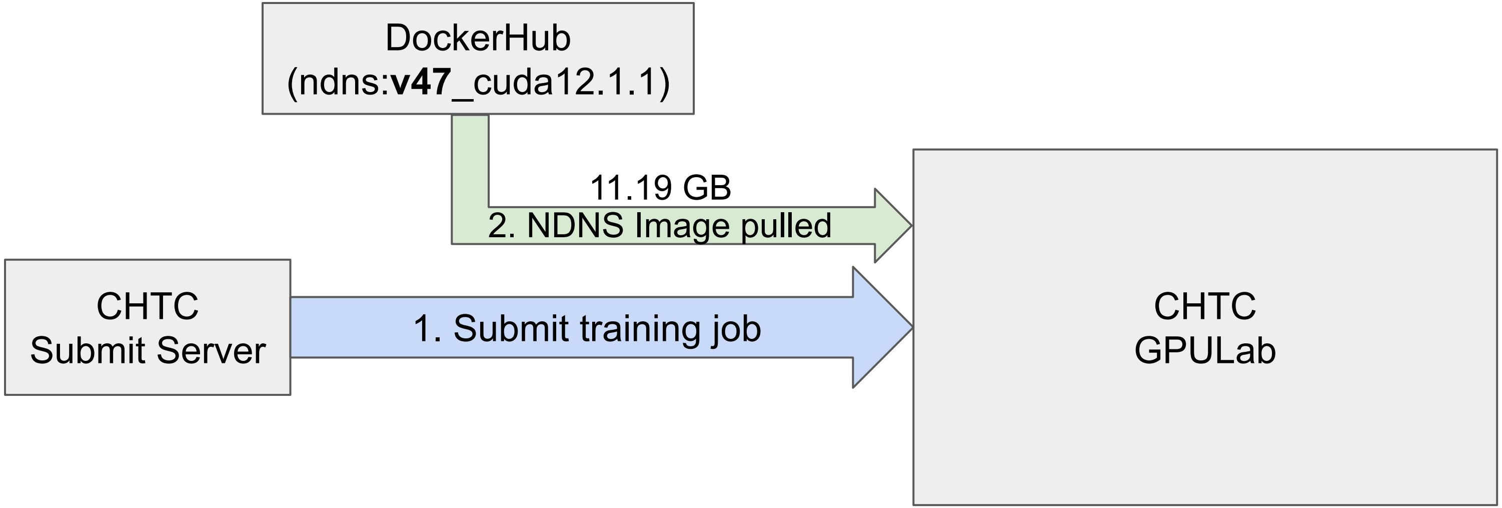
CIPIC dataset allowed us to analyze 1250 possible sound source orientations

* Adapted from: Bear, Mark F. Neuroscience : Exploring the Brain — Fourth edition.

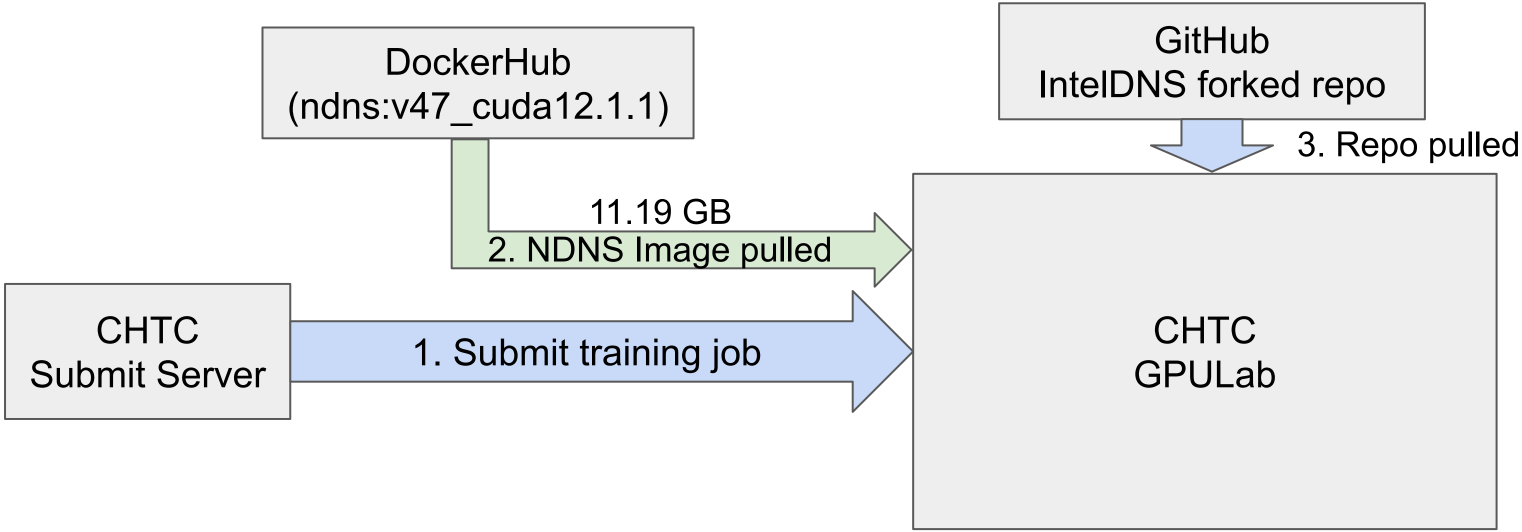
Developing the GPU Workflow (1)



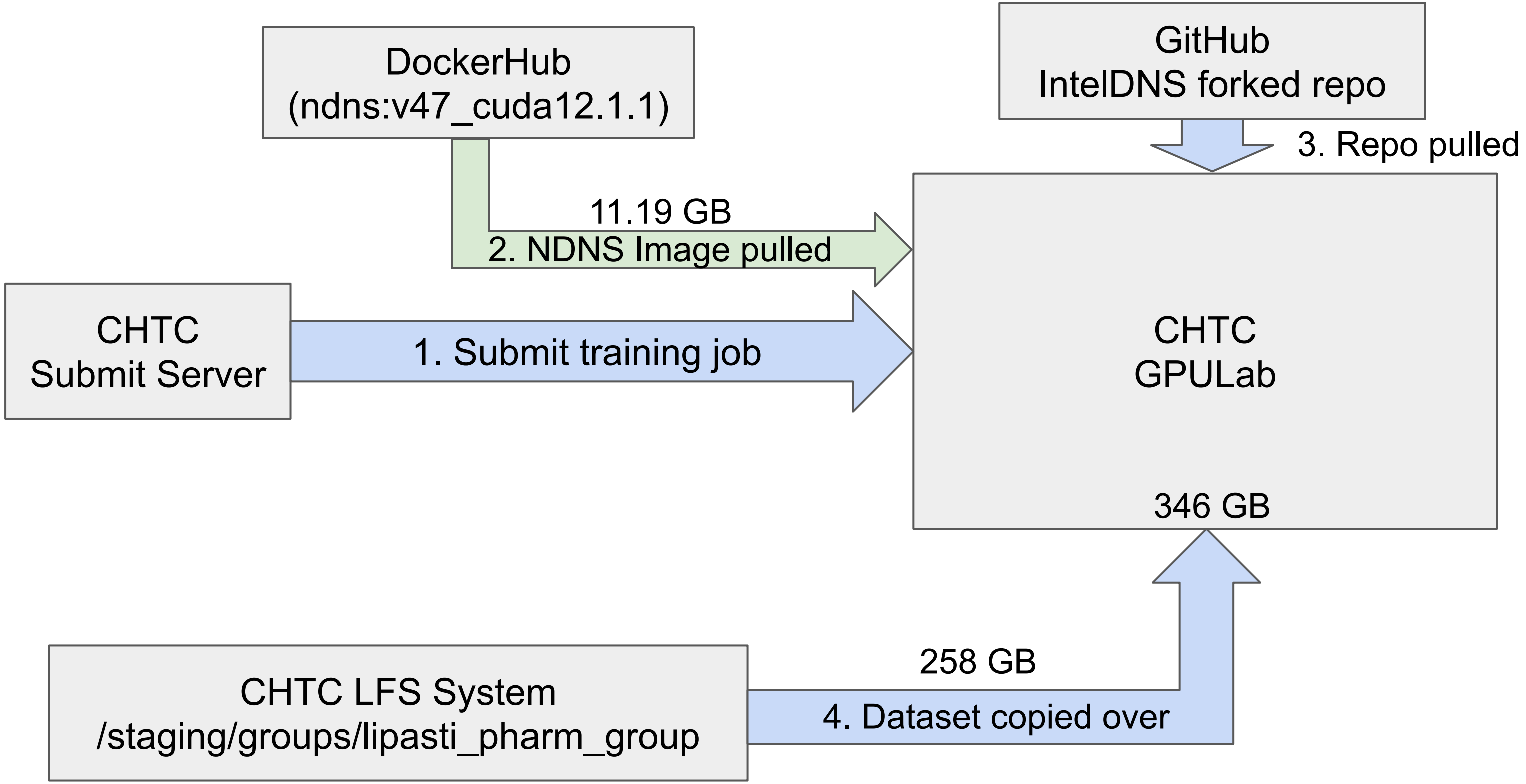
Developing the GPU Workflow (2)



Developing the GPU Workflow (3)

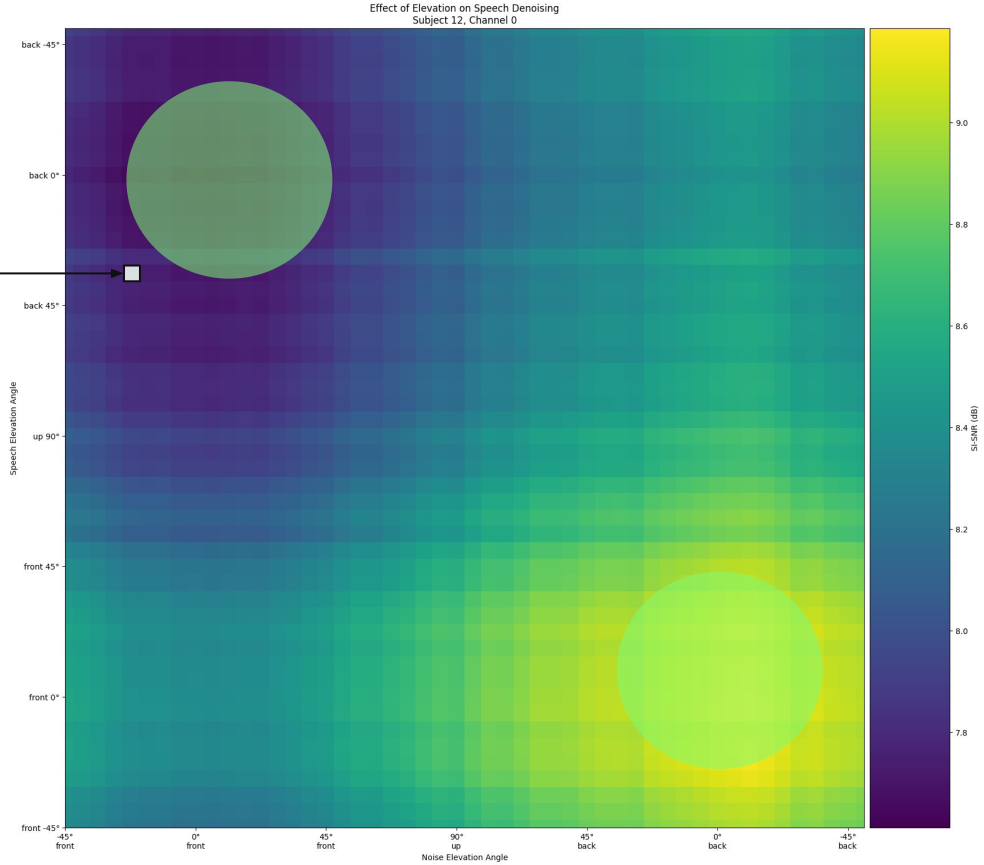


Developing the GPU Workflow (4)



Pinna Results (Baseline)

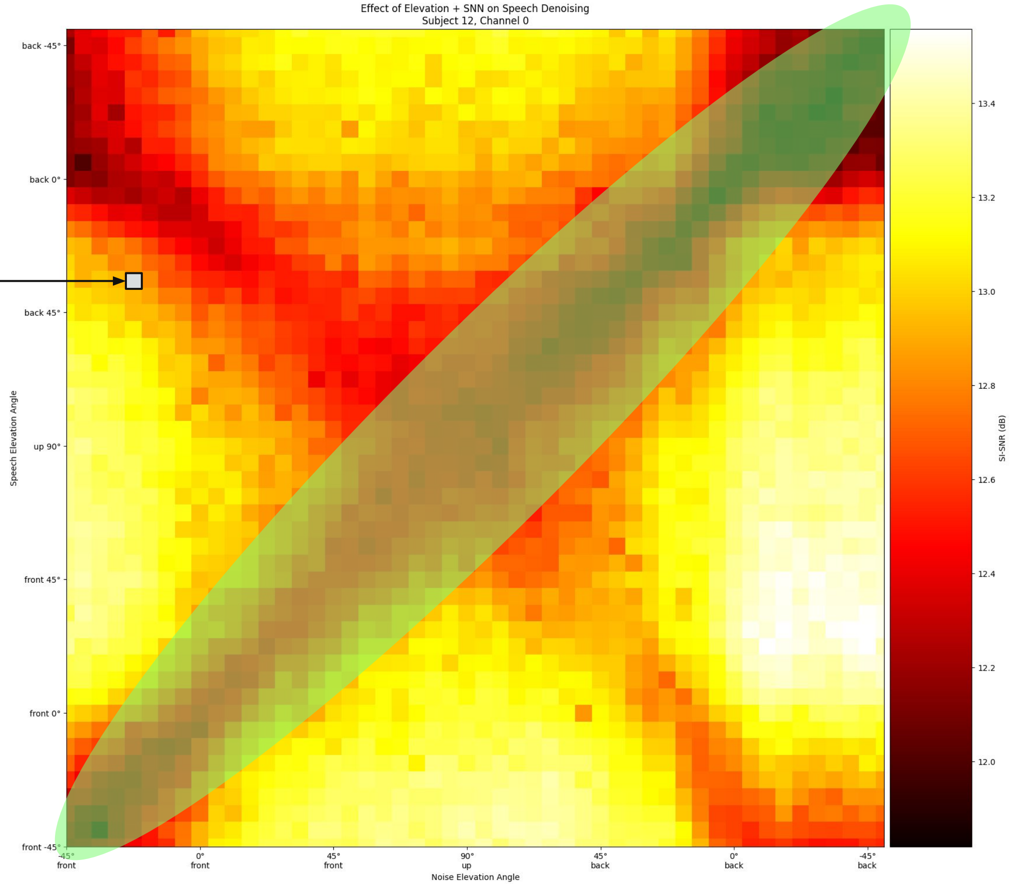
Each trial took ~ 1 hour



Macro scale patterns impossible to view without the computing scale that CHTC provides

SNN Results (Development)

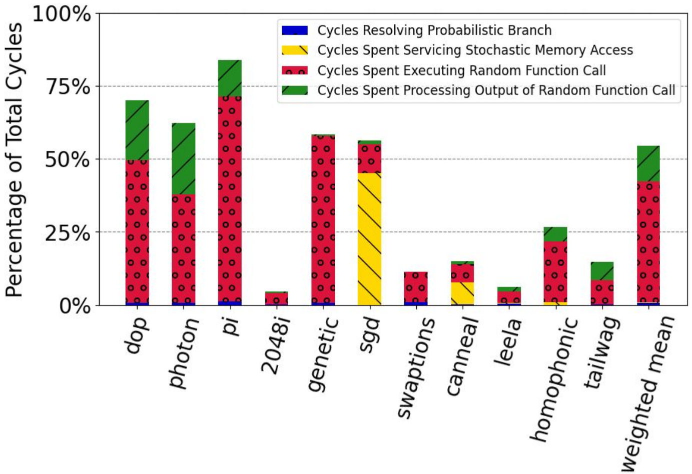
Each trial took ~ 4 hours



Macro scale patterns impossible to view without the computing scale that CHTC provides

Bottom-up: Neuromorphic Workloads are Stochastic

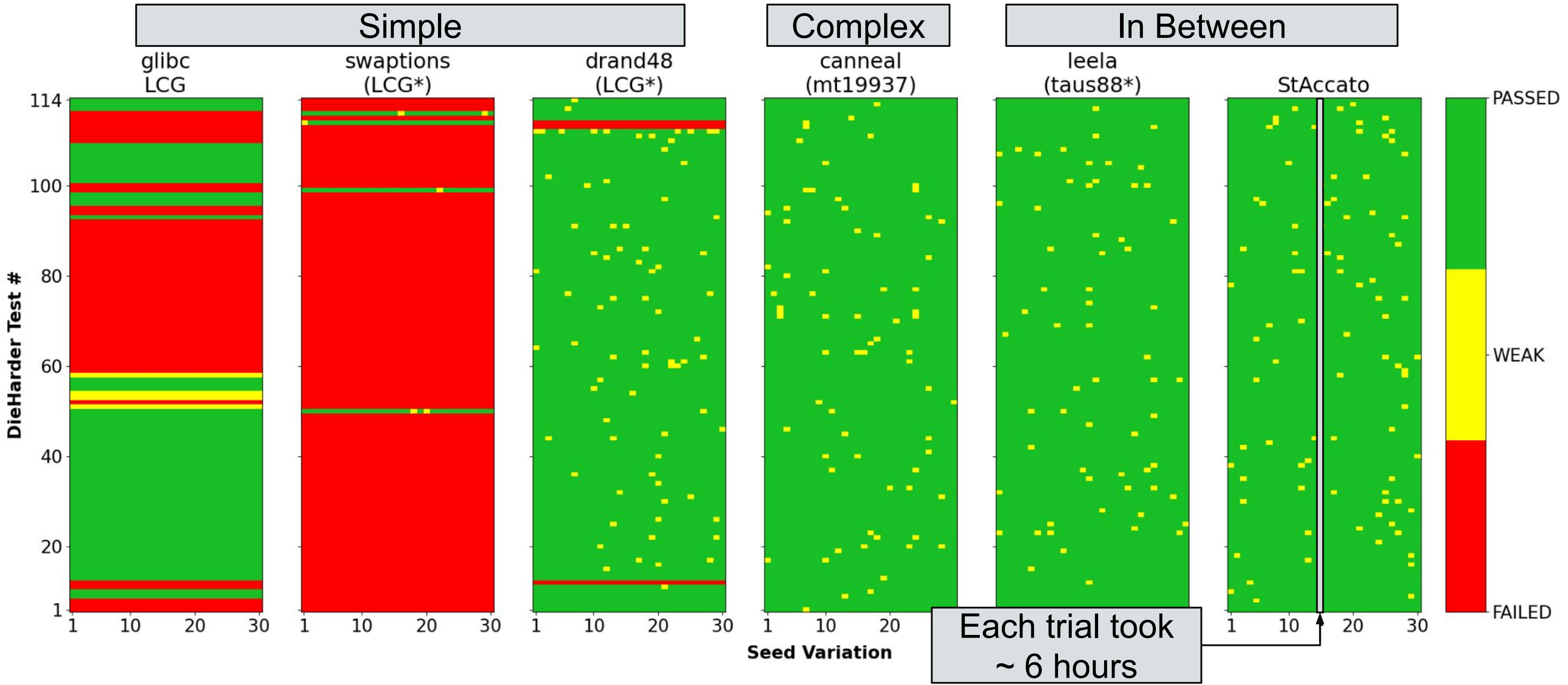
- Key Insight: random number generation and downstream, dependent operations are expensive



- To develop StAccato, a hardware ***acc***elerator for ***st***ochastic workloads, how can we compare RNG quality?

Dieharder

- Great, StAccato is better than simple RNG, but by how much?



Scaling the Problem

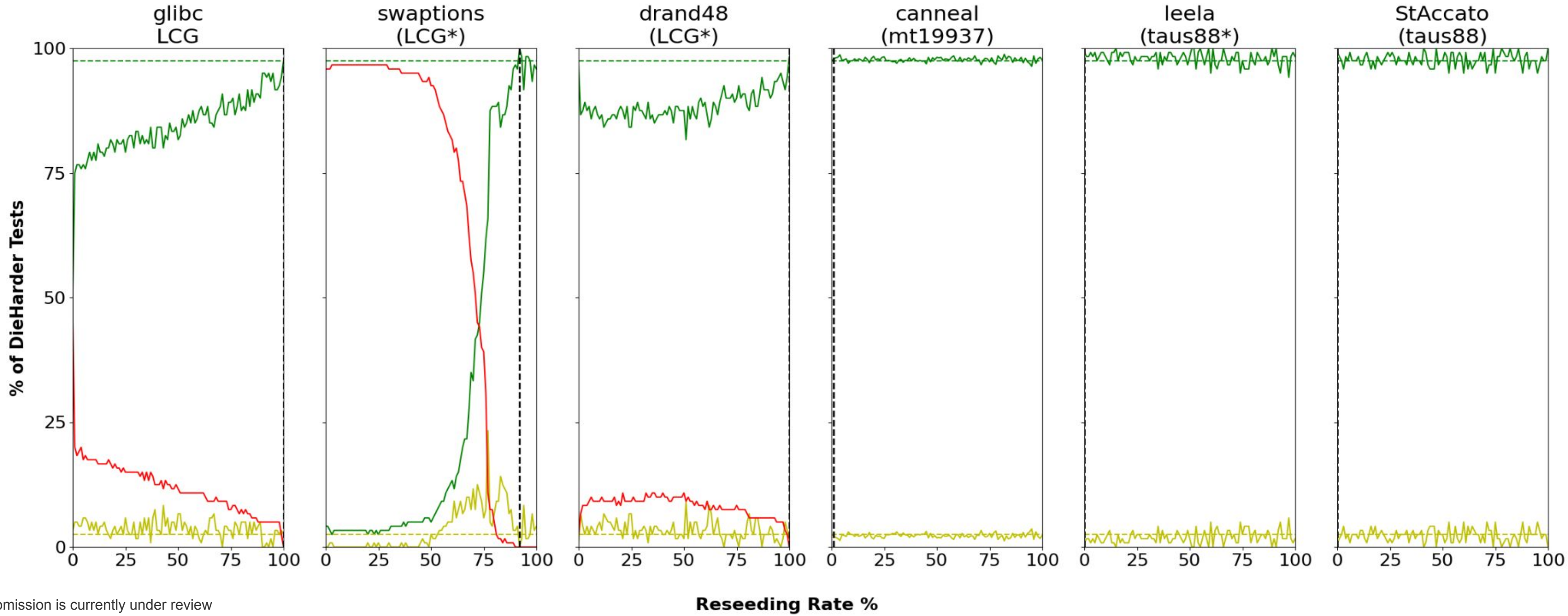
- Throughput problem demanding large amounts of CPU hours, but minimal restrictions:
  - Lightweight docker image with Dieharder installed
  - Single requested CPU, 512MB memory, 1MB storage
  - ~need CPUs past 2011

- Workload is ideal for CHTC's ~40K CPU cores
  - Problem well defined in a single 34 line submit file (+config list)
  - Launched 6 RNGs x 5 trials per rate x 100 different rates from
    - ~2.1 compute years completed in ~3 weeks!

Comparative Dieharder Results

- StAccato is as good as the best of them!

The timeliness of this analysis was only possible via CHTC



* Submission is currently under review

Current Outlook and Next Steps

- Current Outlook:
  - CHTC fairly easy to use and very flexible for a wide variety of studies
  - Minor pain points using the GPU system
    - simultaneous profiling during GPU sim (resolved in an update)
    - non-deterministic crashes (only in < 6% of runs)
    - runtime variability for repeat tasks
- Next CHTC features to explore
  - Checkpointing to support long running GPU sims (> a week)
  - Thorough sweep of model hyper-parameters (width, depth, fft bins, etc.)
  - Better coordination of job resources in allocation request and during runtime
- Desirable features from CHTC in the future
  - Vendor variety (AMD GPUs)

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Thank You!