RAISE
Center of Excellence

Experience running workloads on CPU+GPU at HPC

HEPscore 2022 Fall Workshop

Motivation

Computing environments have evolved rapidly since the time of HEPSpec06:

2006 landscape:

- Dual core processors
- 32-bit instruction set
- Few gigs memory
- Homogenous

2022 landscape:

- 64 cores (and HyperThreading…)
- 64-bit, diverse instruction sets
- TBs memory
- Specialized accelerators

x86 Benchmarking is well-covered by HEP experiment code bases

Differences in HPC

High-performance computing centers offer immense compute potential, but adoption challenges exist:

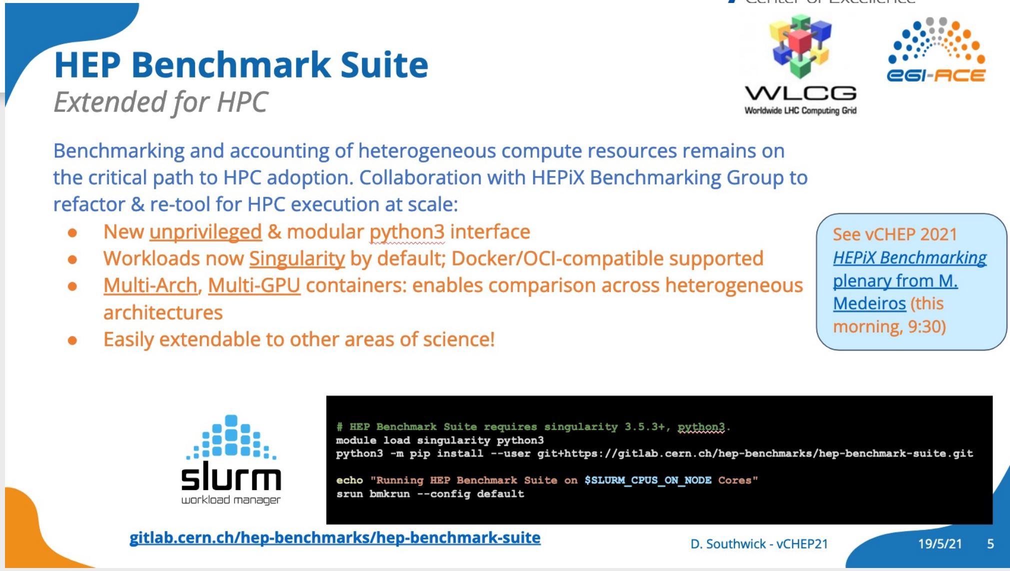
- Completely unprivileged environment: ”BYOE” – bring your own environment - with containers, monitoring, and reporting
- ”Growing pains” – I/O bound performance requires careful consideration when scaling jobs on **shared** storage; ingress & egress
- Lack of site information about high-throughput computing (HTC) support

Much development over the past years within the Benchmarking WG to address these challenges and enable execution at scale!

Successes at HPC centers

HEPscore (executed by the HEP-Benchmark-Suite) has already been used for large scale deployments and studies at HPC sites:

- Initial experiences from vCHEP'21
- 200,000-core campaign with Run-2 production WLs
- Scale studies of new/upcoming AMD cpus



HEP Benchmark Suite

Extended for HPC

Benchmarking and accounting of heterogeneous compute resources remains on the critical path to HPC adoption. Collaboration with HEPiX Benchmarking Group to refactor & re-tool for HPC execution at scale:

- New unprivileged & modular python3 interface
- Workloads now Singularity by default; Docker/OCI-compatible supported
- Multi-Arch, Multi-GPU containers: enables comparison across heterogeneous architectures
- Easily extendable to other areas of science!

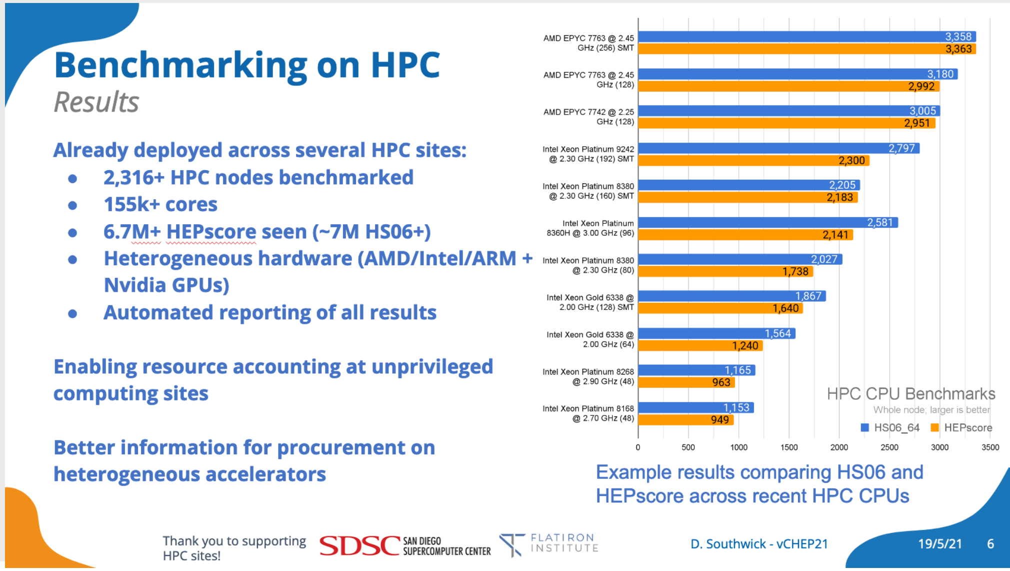
See vCHEP 2021 *HEPiX Benchmarking plenary from M. Medeiros* (this morning, 9:30)

```
# HEP Benchmark Suite requires singularity 3.5.3+, python3.
module load singularity python3
python3 -m pip install --user git+https://gitlab.cern.ch/hep-benchmarks/hep-benchmark-suite.git

echo "Running HEP Benchmark Suite on $SLURM_CPUS_ON_NODE Cores"
srun bmkrun --config default
```

gitlab.cern.ch/hep-benchmarks/hep-benchmark-suite

D. Southwick - vCHEP21 19/5/21 5



Benchmarking on HPC

Results

Already deployed across several HPC sites:

- **2,316+ HPC nodes benchmarked**
- **155k+ cores**
- **6.7M+ HEPscore seen (~7M HS06+)**
- **Heterogeneous hardware (AMD/Intel/ARM + Nvidia GPUs)**
- **Automated reporting of all results**

Enabling resource accounting at unprivileged computing sites

Better information for procurement on heterogeneous accelerators

Example results comparing HS06 and HEPscore across recent HPC CPUs

Thank you to supporting HPC sites!

D. Southwick - vCHEP21 19/5/21 6

See vCHEP'21 presentation for full details

What’s new?

First look of run3 workloads - many with heterogenous architectures!

- First ARM, IBM POWER, GPU development workloads
- GPU *vs* CPU *vs* GPU+CPU benchmarking studies
- Heterogenous partition studies (ARM+GPU, POWER+GPU, etc)
- ML / AI workload development (MPI scaled to ~200 GPUs)

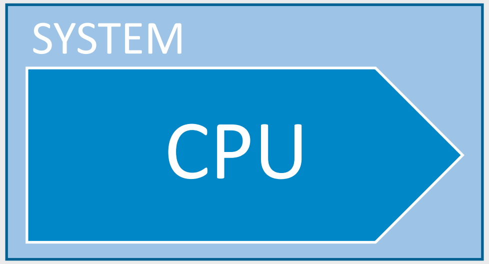
Quality-of-Life updates:

- Batch uploading (post-run: supports “secure” worker nodes)
- GPU / accelerator meta-data inclusion
- CVMFS-attached benchmarking campaigns

Thank you to all partner HPC sites enabling this work!

Workloads

CPU Benchmarks



SYSTEM

CPU

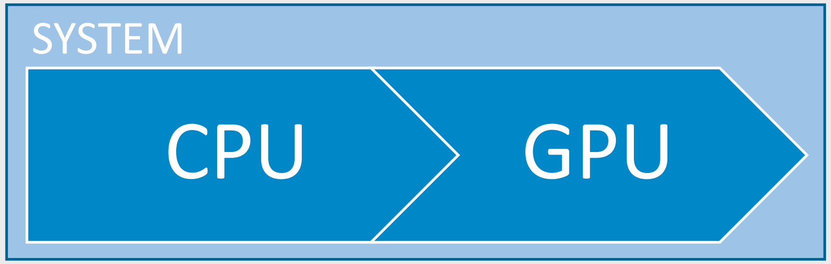
Experiments have been hard at work exploiting additional instruction sets outside of traditional x86:

- ARM: development workloads available via HEPscore
- POWER: candidate workloads in testing
- Open: (eg OpenCL, Python-based, etc) – available via HEPscore

All these workloads follow the same familiar recipe: *F (events/sec.)*

Additional workloads expected as development progresses

Heterogeneous compute environments



Unclear what GPU benchmarks will be adopted…
…perhaps this void might be filled?

Considerable percentages of site total computing power increasingly reside in GPUs. HEP workloads with "simple" kernels (*embarassingly parallel)* can profit by orders of magnitude – HEPscore provides workloads that run on both:

- MadGraph – (one of) main generator for HL-LHC
- CMS-HLT – GPU accelerated HLT
- ML ParticleFlow – evolution of AI processing

Typical HPC single node resources:

2x AMD EPYC:	256 threads
4x Nvidia V100:	20,480 cuda cores
4x Nvidia A100:	27,648 cuda cores
4x Nvidia H100:	*67,584 cuda cores**

* https://www.nvidia.com/en-us/data-center

GPU workload performance

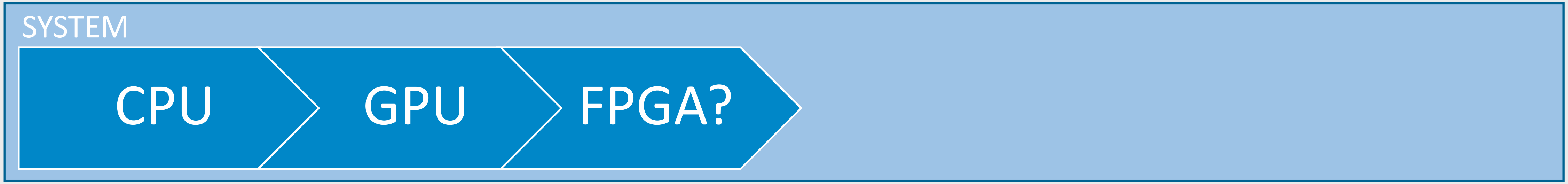
Preliminary testing on HPC enables direct comparison of same codebase and same hardware:

- Xeon Gold 6148 @ 2.4Ghz, Nvidia V100

Workload	CPU only	GPU only	Speedup	Time(CPU)	Time(GPU)
MadGraph5	0.026(float)	0.744	28x	29m 8s	11m 8s
CMS-HLT	525	9,450	18x	23m 9s	17m 15s
ML particle flow (epoch time)	659s	138s *1 GPU	4.8x	33m 36s	8m 29s

Non-production development values
Results likely to improve*

…and more to come!



Increasingly specialized compute partitions are available at HPC sites today:

Several sites already offer FPGA partitions, potentially enabling future workloads to achieve event throughput near clock speed. Future accelerators can be tested today with HEPscore workloads based on OpenCL or Python – enabling them to run on *any* hardware that supports the frameworks

- Simpletrack
- ML ParticleFlow

Future outlook

Lots of development this past year accelerated by HPC, and certainly momentum will only accelerate!

- Benchmark I/O performance, scaling benchmark for HPC
- First AMD GPU partitions coming online ~October'22 (LUMI-G)
- ARM partitions coming later this fall
- openMPI workloads (ML, distributed jobs)

Stay tuned for dedicated talks on many of the workloads mentioned in this presentation drive. enable. innovate.

SDSC SAN DIEGO SUPERCOMPUTER CENTER

FLATIRON INSTITUTE

JÜLICH Forschungszentrum

RAISE Center of Excellence

The CoE RAISE project has received funding from
the European Union's Horizon 2020 –
Research and Innovation Framework Programme
H2020-INFRAEDI-2019-1 under grant agreement no. 951733

Follow us:

Backup

Benchmarking HPC capabilities that are *not* compute (eg shared storage throughput, WAN, etc)

- I/O benchmark based on HEP workload I/O patterns developed
- ”Rapid” benchmark for estimating scaling capabilities of I/O bound jobs
- Ideal for opportunistic computing
- Looking for collaborators to further development!