Hewlett Packard
Enterprise

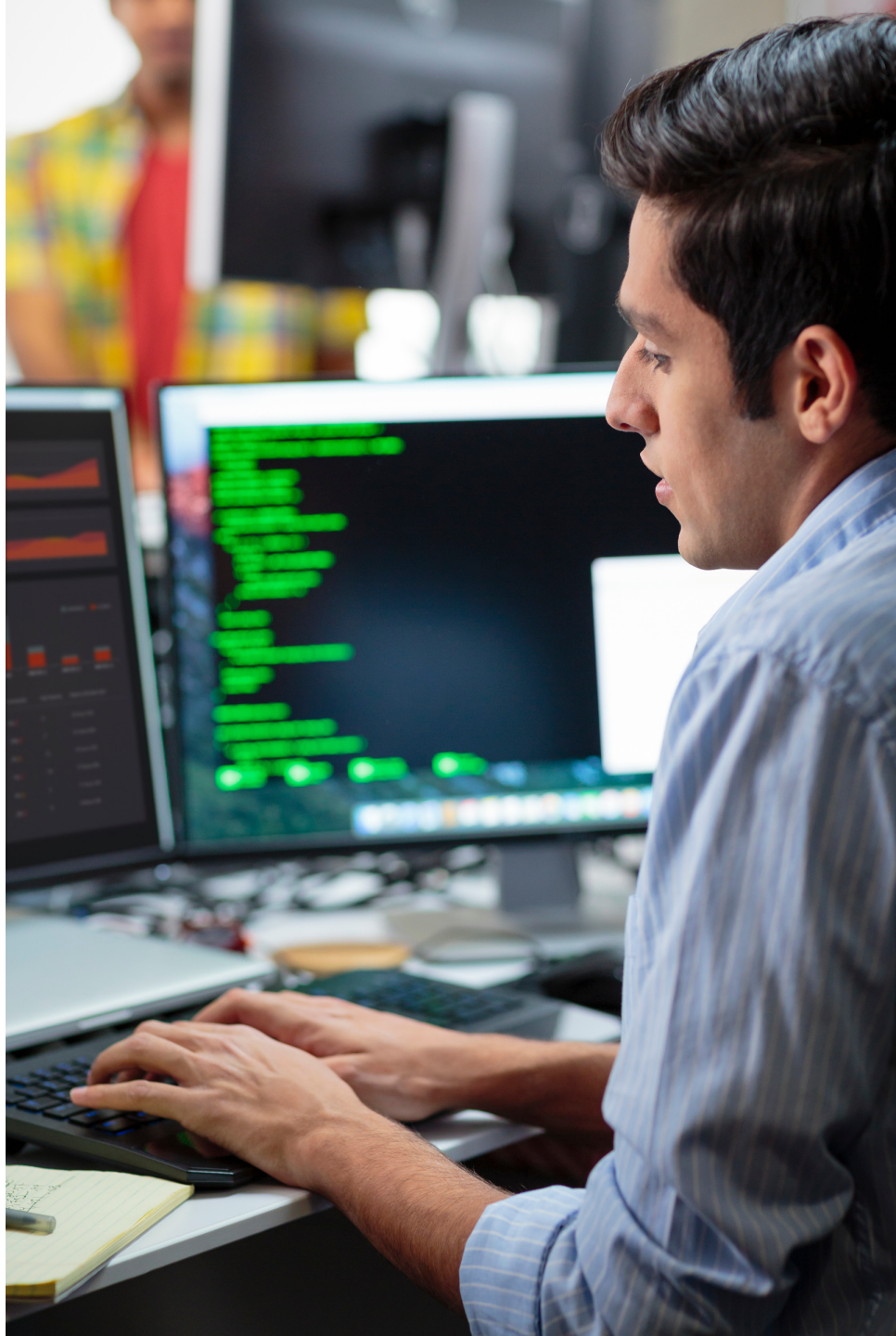
HPE CRAY PROGRAMMING ENVIRONMENT UPDATE

Barbara Chapman, Distinguished Technologist

May 9, 2023

AGENDA

- Introduction to HPE Cray Programming Environment (CPE)
- Recent Features
- What's Next?
- Future Directions



HPE CRAY PROGRAMMING ENVIRONMENT

Essential toolset for HPC organizations developing HPC code in-house

Fully integrated software suite with compilers, tools, and libraries designed to increase programmer productivity, application scalability, and performance.

Complete toolchain
For the whole application development process.

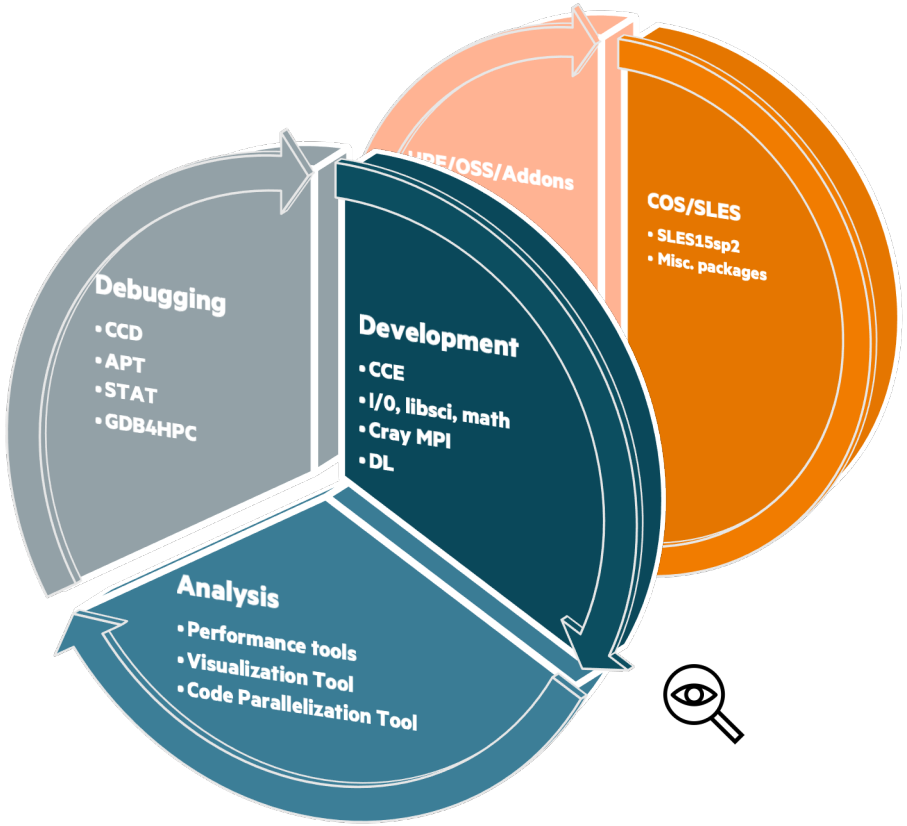
Holistic solution
Unlike processor-specific tools, the suite enables software development for the full system (including CPUs, GPUs and interconnect) for the best performance.

Programmability
Offering users intuitive behavior, automation of tasks and best performance for their applications with little effort.

Scalability
Improving performance of applications on systems of any size—up to Exascale deployments.

Complete Support
HPE Pointnext Services support the whole suite, not just the tools we developed.

From HPC experts for HPC experts
Developed for over 30 years in close interaction and contributions from our users.



HPE CRAY PROGRAMMING ENVIRONMENT

Comprehensive set of tools for developing, porting, debugging, and tuning of HPC applications on HPE & HPE Cray systems

Development

Programming Models

- HPE Cray MPI
- SHMEM
- OpenMP | OpenACC (Fortran)
- AMD ROCm HIP | NVIDIA CUDA
- UPC | Fortran co-arrays
- Global Arrays
- Chapel

Programming Environments

- Compiling Environment
  - GNU
  - Intel Programming Environment
  - AMD Programming Environment
  - NVIDIA HPC SDK

Programming Languages

- C
- C++
- Fortran
- Python
- R

Optimized Libraries

- LibSci (BLAS)
- LAPACK & ScaLAPACK
- LibSci_ACC
- Iterative Refinement Toolkit
- FFTW

I/O Libraries

- NetCDF
- HDF5

Debugging

Comparative Debugger
Compare two versions of an application

GDB for HPC
Parallelized gdb for HPC

Valgrind for HPC
Memory debugging at scale

Stack Trace Analysis Tool
Stack tracing at scale

Sanitizers for HPC
Detect memory and thread errors at scale

Tool for Abnormal Termination Processing
Manage core files at scale

- TotalView
- DDT

Performance Analysis & Optimization

Performance Analysis Tool (PAT)
Whole program performance analysis, exposing wide set of indicators, identifying bottlenecks and automatically generating suggestions to improve performance.

Visualization Tool
Complements text reports with summary of performance data in graphs and charts, allowing users to drill down and resolve issues

Code Parallelization Assistant
Reveal hidden potential of an application via code restructuring

Setup & Runtime

Environment Setup

Tool Enablement
(for Spack, CMake, EasyBuild, etc)

Modules/Lmod

Supported systems:

- HPE Cray EX
- HPE Apollo
- HPE Cray XD
- HPE ProLiant

HPE—authored | HPE Added-value to 3rd party | 3rd party

MAJOR ACCOMPLISHMENTS IN 2022

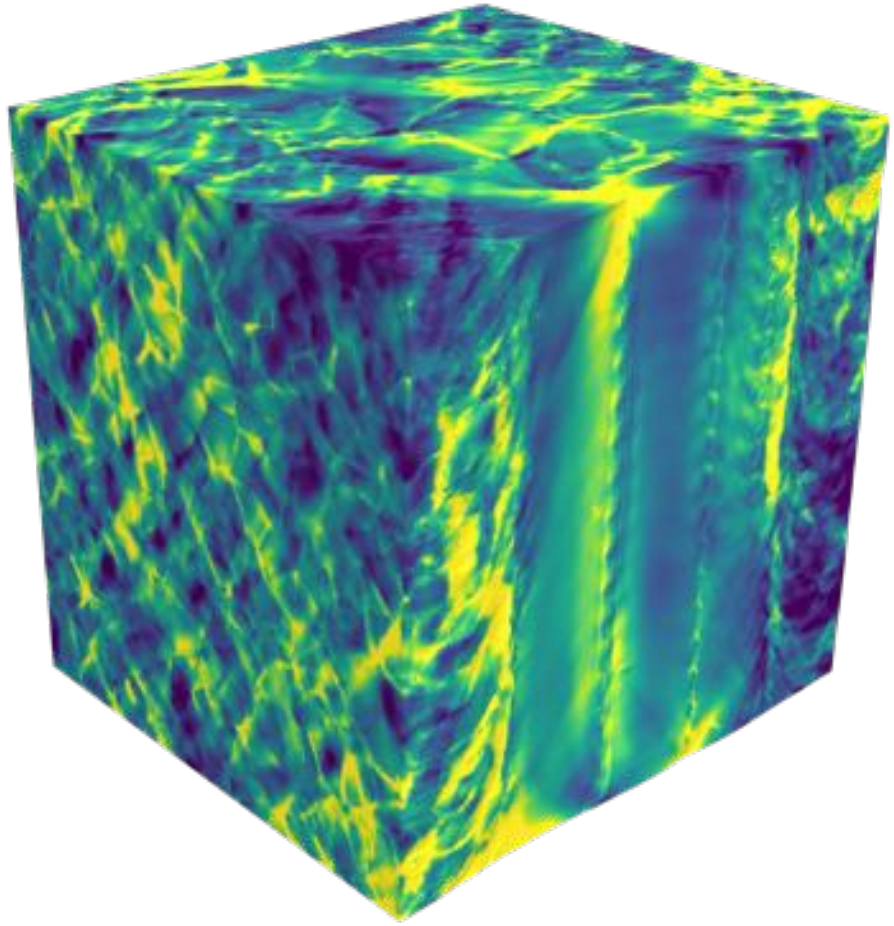
- HPE Cray MPI was used to achieve 1.102 Exaflop/s on ORNL's Frontier supercomputer!
  - It's also the primary MPI implementation on several other major HPE EX systems: EuroHPC-CSC (LUMI, #3 on TOP500), NERSC (#8), CINES (#11), Pawsey (#15), ExxonMobil (#16)
  - Frontier TDS "Crusher" system claimed the #1 spot for the GREEN500 list (June 2022) also relies on HPE Cray MPI

INTEGRATED TOOLSET: SUPPORTING SCIENCE APPLICATIONS ON FRONTIER

ExaConstit: Open-source crystal-plasticity FEM code

- Exascale-ready open-source simulation
  - Part of Integrated Platform for Additive Manufacturing Simulation (IPAMS) suite
  - Directly incorporates microstructure evolution and effects of microstructure within AM process simulation.
- Developed in ECP project

- Point-to-point communication was unexpectedly slow
- *pat_report* showed excessive time in *hipMemcpy*
  - Changed to GPU-aware MPI, with little improvement
- *pat_report -O acc_time* showed the real culprit!
  - *PAT_RT_PERFCTR=_busy_0* showed the suspect had *MemUnitBusy* 94.1%
- Tuning (reforming?) the culprit improved overall runtime by 2.5x



WHAT'S NEW IN SUPPORTED PLATFORMS?

Target Architectures

- AMD Genoa, Milan-X; Intel Sapphire Rapids; Nvidia Grace (ARM)
- GPUs: AMD MI250X, Nvidia Ampere (A100)

WHAT'S NEW IN LANGUAGES AND COMPILERS?

Parallel Programming Models and Language Features

- Full OpenMP 5.0 support along with major progress towards OpenMP 5.2; working toward OpenACC 3.3
- C/C++ continued tracking of LLVM - CCE 16.0 with the integration of latest stable LLVM 16 release this month
- OpenMP interoperability support for GNU (CPU only) & Clang-based compilers (CPU, GPU)
- Starting Fortran 2023 language features
- Chapel 1.30 compiler generates NVIDIA and AMD GPU kernels for some foreach/forall loops *
- CPU sanitizers for C, C++, & Fortran
- DWARF5 Support

Performance Optimizations

- Optimized offload kernel launch and loop scheduling performance
- Relaxed locking and increased concurrency for multi-threaded use of GPU
- Tuned loop unrolling heuristic for AMD GPU targets
- Loop restructuring (not just GPU)
- Improved "omp parallel" performance
- SVE Support (Nvidia Grace)

* https://chapel-lang.org/docs/technotes/gpu.html

WHAT'S NEW WITH SCIENTIFIC/3RD-PARTY LIBRARIES?

Scientific Libraries

Cray LibSci

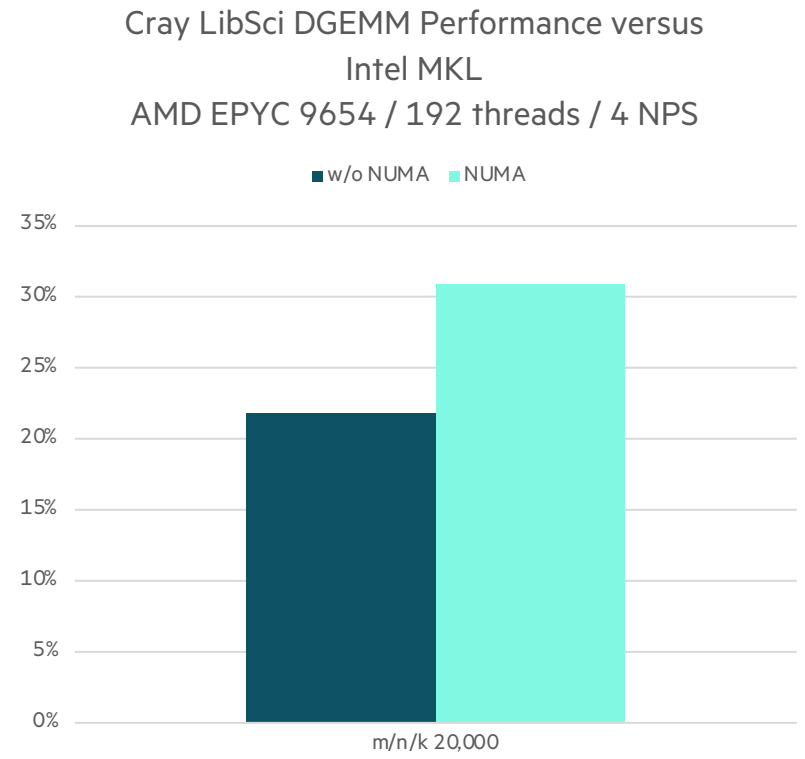
- CPU dense linear algebra routines (full BLAS, LAPACK, ScaLAPACK implementations) optimized for new architectures, especially AMD Genoa

Cray LibSci_ACC

- Additional optimized hybrid CPU-GPU algorithms
- Extensive tuning for new target architectures
- Improved multimode support and tuning for distributed routines

Cray FFTW

- Optimized kernels for AMD Genoa CPU targets



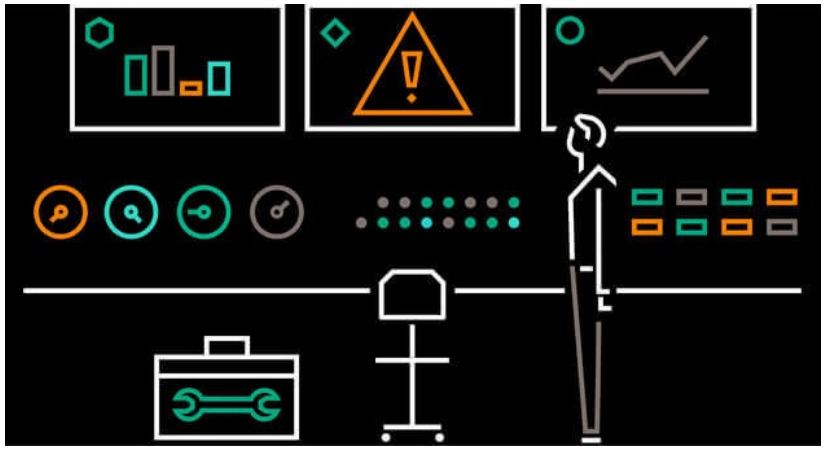
Data Libraries

HDF5, NetCDF, Parallel NetCDF built for AMD Genoa CPU and a range of different compilers

WHAT'S NEW IN DEBUGGING TOOLS?

Gdb4hpc features

- C++: complex templates, anonymous classes, anonymous namespace
- Conditional breakpoints and breakpoint counters
- Remote shell command
- Pipe the result of any command to a file or application
- Support for Kokkos::View and RAJA::View types
- Python mode; view program data with numpy or python libraries
- Run command to relaunch keeping breakpoints
- Direct access to Cuda commands
- Support for non MPI programs
- Extended tutorials

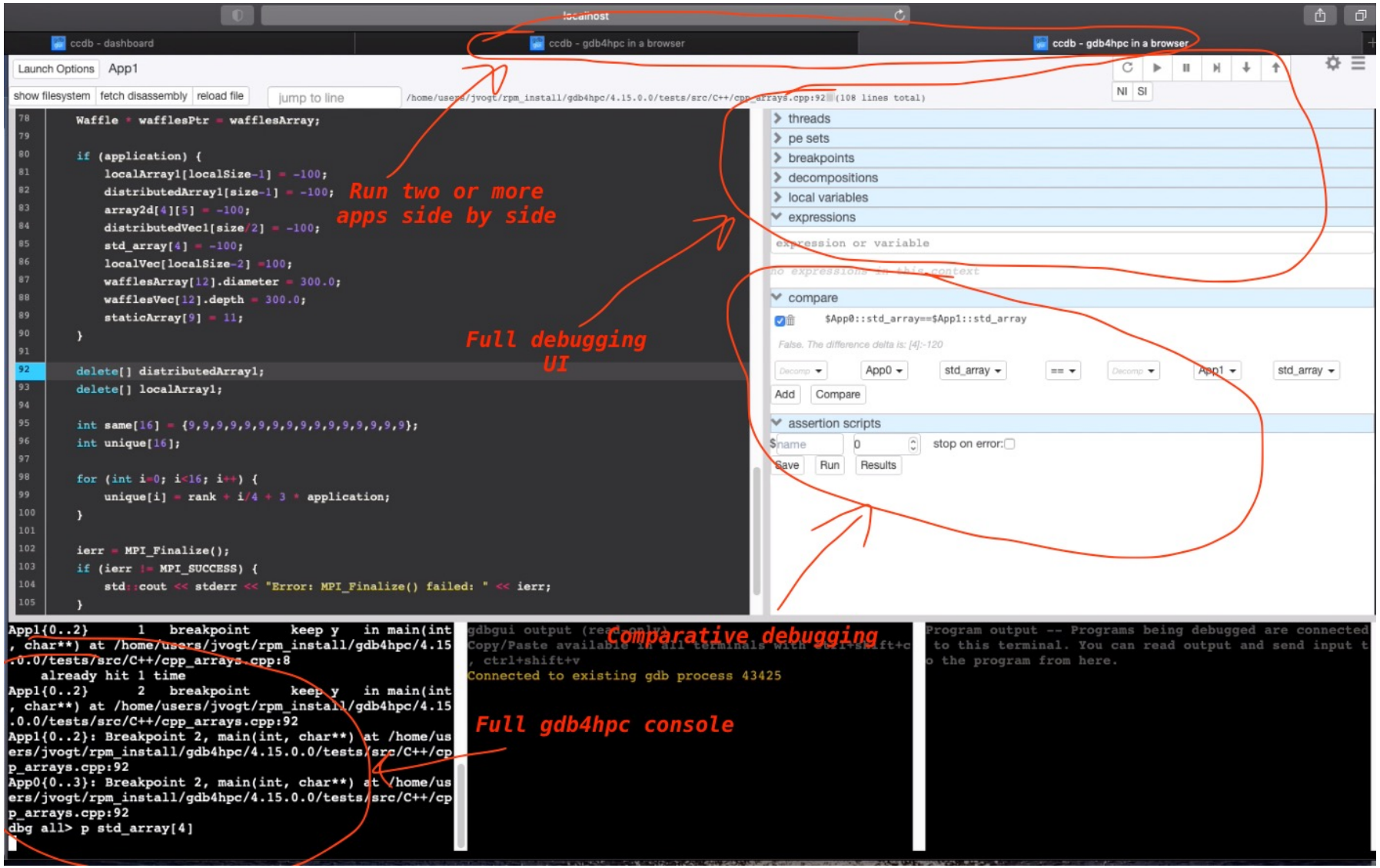


Other tool enhancements

- Ccdb: 5.0 with browser based UI
- ATP: slurm plugin now works in user space; no sysadmin work required
- Sanitizers4hpc: Cuda-memcheck
- Valgrind4hpc: enabled interactive debugging with gdb4hpc when an error is detected

CCDB 5.0

Run a full featured debugging GUI on your laptop even with limited bandwith



WHAT'S NEW IN PERFORMANCE TOOLS?

Performance Analysis Tool (PAT)

Improved support for GPUs

- Nvidia A100 performance counters
- Revised Nvidia V100 & A100 counter groups
- Specialized setup mechanics to support profiling of multiprocess applications

Push/Pop Region API

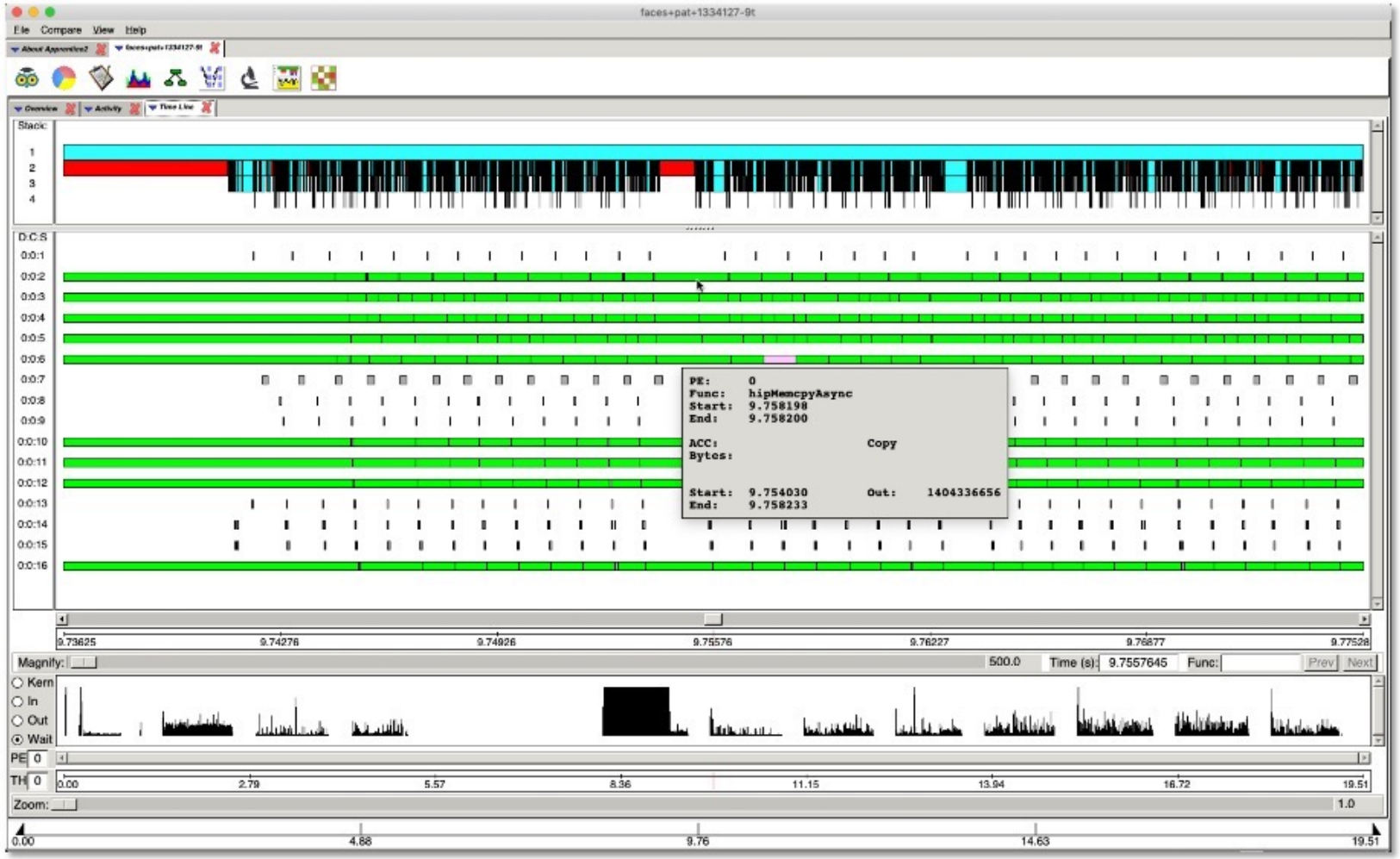
- Improved support for Kokkos/Raja codes
- Made push/pop region API available in the default callstack mode
- This improves our Caliper support as well

Added detailed power reporting, breaking down power usage per node by CPU, GPU, memory

Python Profiling

- Reduced runtime overhead

CORRELATE ACTIVITY OF HOST AND DEVICES



WHAT'S NEW WITH MESSAGE PASSING?

New HPE Cray MPI Features

Cassini Counters and Network Timeout Feature
MPIxlate – Application Binary Interface (ABI) Translator for MPI Programs
Spawn / Connect / Accept Support on Slingshot-11 (Slurm and PBSPro/PALs)
Support for GPU-NIC Async Stream Triggered Point-to-Point Operations
Enhanced Slinghot-11 Traffic Class Interaction with WLMs

Key Optimizations

Collective Performance Improvements for MPI_Igatherv, MPI_Alltoall and MPI_Iallgather/MPI_Iallgatherv
Support for > 2 GB message sizes for MPI_Igather, MPI_Scatter and MPI_Iscatter
GPU Collective Performance Improvements for MPI_Allgather, MPI_Allreduce and MPI_Reduce_scatter_block
Enabled GPU kernel-based optimizations by default for select collectives
Optimized MPI_Graph_create (support for reorder flag)

Workload Management Enhancement

Enhanced Cray MPI Spack support for Nvidia, AMD and Intel compilers in CPE

CASSINI COUNTERS AND NETWORK TIMEOUT FEATURE

Slingshot-11 Network Timeout Events

Network Timeouts (link flaps) cause packets to be automatically re-issued
May cause delays or sub-optimal performance（app dependent）
MPI tracks network timeout events for each MPI job
If network timeouts affected the job, MPI displays:

[MPICH Slingshot Network Summary: 6 network timeouts]

Cassini Counters

MPI collects and summarizes Cassini counter data for each job
Helpful for optimizing MPI communication and debugging performance issues
See MPICH_OFI_CXI_COUNTER_REPORT in man page for more options

```
MPICH Slingshot CXI Counter Summary:
Counter                                Samples          Min          (/s)         Mean          (/s)          Max          (/s)
atu_cache_evictions                      34000       223236        1099.7      1298417        6396.1      4812004       23704.5
atu_cache_hit_base_page_size_0           34000           14           0.1          585           2.9         3225          15.9
atu_cache_hit_derivative1_page_size_0    34000   3567598279  17574375.8   3580464549  17637756.4   3586507556   17667524.9
lpe_net_match_priority_0                 34000     14292803      70407.9     16977380       83632.4     18460483       90938.3
lpe_net_match_overflow_0                 34000      3571835      17595.2      5055416       24903.5      7739576       38126.0
lpe_net_match_request_0                  34000           15           0.1          708           3.5         2861          14.1
lpe_rndzv_puts_0                         34000     11014704      54259.6     11014704       54259.6     11014704       54259.6
lpe_rndzv_puts_offloaded_0               34000      5831837      28728.3      8447950       41615.5     10441812       51437.5
hni_rx_paused_0                          34000     17771789      87545.8   3166422790    15598141.8  12685217652    62488756.9
hni_rx_paused_1                          34000     32928226     162208.0    122307495      602500.0   1165354793     5740664.0
hni_tx_paused_0                          34000         2602         12.8       512934        2526.8      3246672       15993.5
```

MPI TRANSLATION TOOL IN HPE CRAY PROGRAMMING ENVIRONMENT

MPIxlate is a new feature for HPE Cray MPI released in November 2022

Open MPI and HPE MPI are not ABI-compatible with MPICH-based MPIs

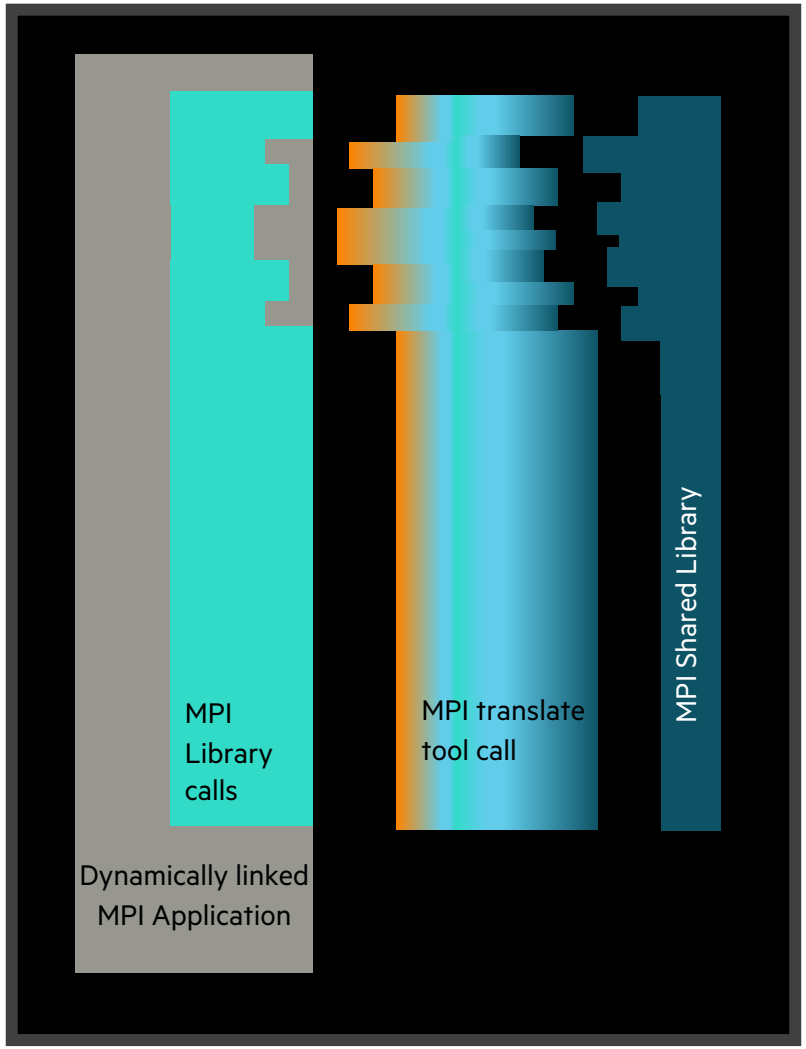
MPIxlate transparently translates Open MPI ABI and HPE MPI ABI to MPICH ABI at runtime

Enables MPI applications compiled with Open MPI or HPE MPI to run with HPE Cray MPI

Specifications:

- Supports C and C++ Languages
- Supports CPUs and GPUs

Documentation (available via Quick Help, Man page & Command line)



WHAT'S NEXT FOR CPE?

Hardware support

AMD APU (MI300 A)
Grace+Hopper (and H100 standalone)
Intel GPU Support (MPI)
Next-gen Slingshot

Flexible deployment

Changes in module environment
Standalone components
Development and runtime containers
Spack binary cache support

DevOps and user support enhancement

Better and more accessible documentation
Online tutorials
Improvements to testing procedures

WHAT'S NEXT FOR CPE?

CCE compilers and CSML

OpenACC 3.3 and OpenMP 5.2 features
GPU-enabled DO CONCURRENT
Generic GPU-enabled DWARF
Further development of LibSci_ACC

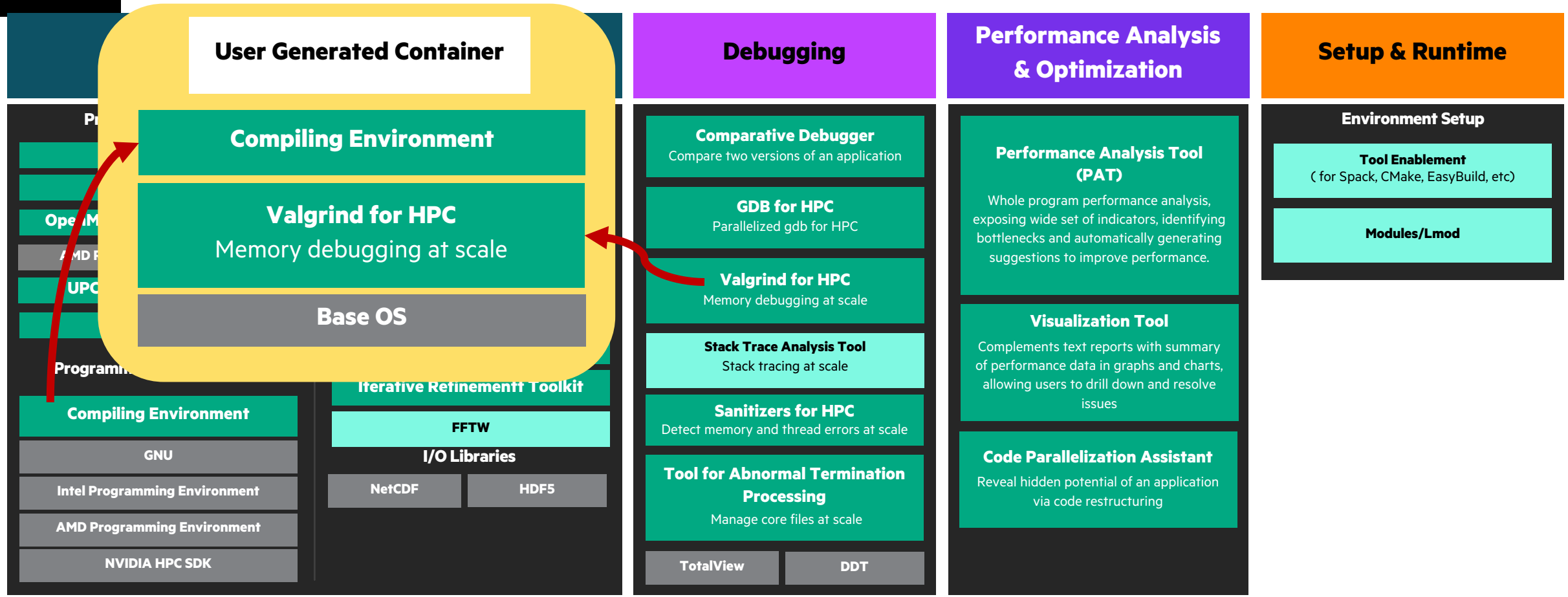
MPI

Support for the MPI 4.0 Standard
Support for Hardware-accelerated Collectives
Select Non-blocking Collectives using Cassini Triggered Ops
Thread-hot RMA Optimizations
MPI RMA Communication using GPU-NIC Async Stream Triggered Ops

Debuggers, Performance Analysis and Optimization Tools

Improvements to user interfaces
GPU PC sampling support for intra-kernel profiling

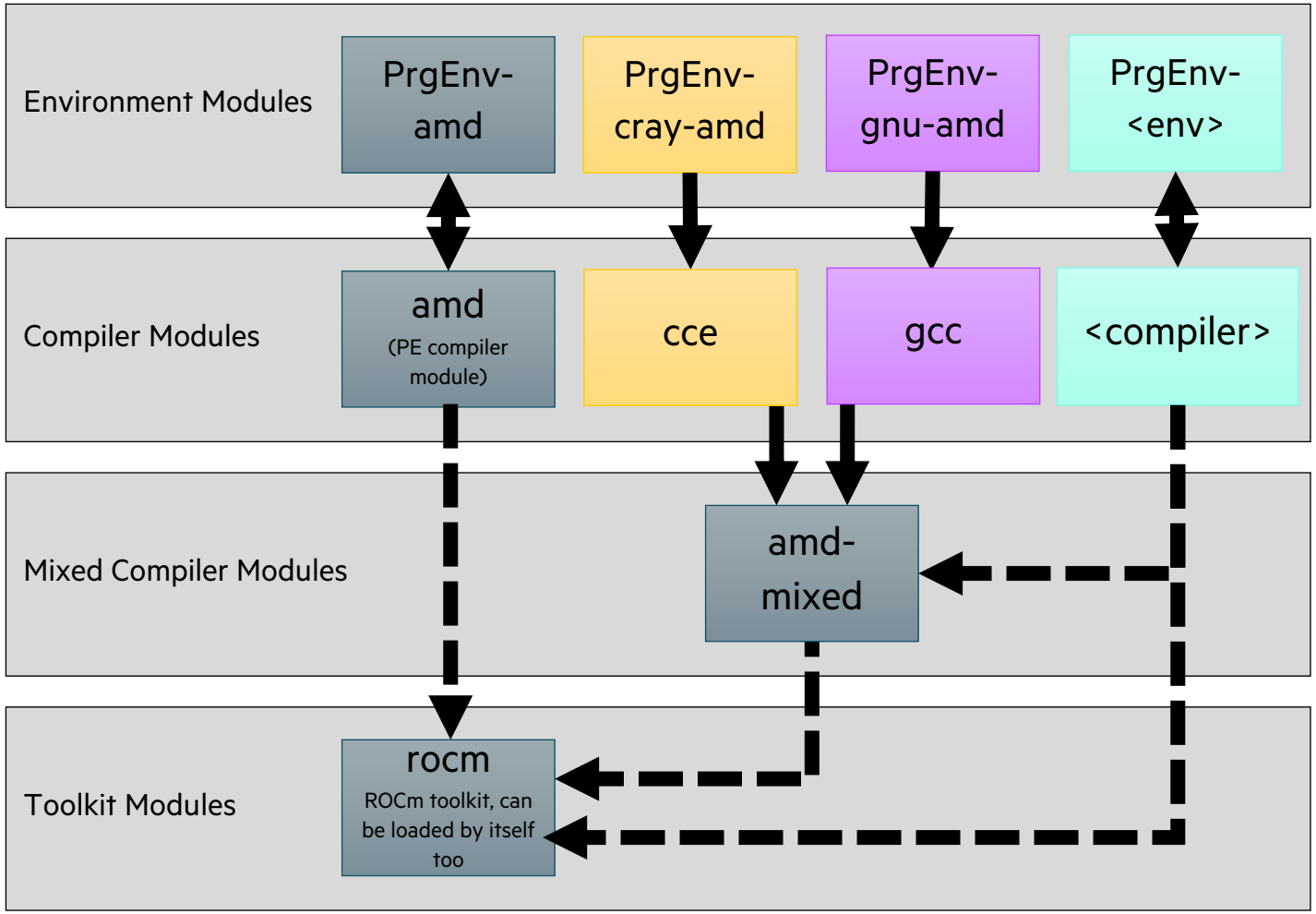
HPE CONTAINERIZED CRAY PROGRAMMING ENVIRONMENT



HPE—authored | HPE Added-value to 3rd party | 3rd party

MODULES

- Modules ensure that appropriate versions of libraries and tools are loaded
  - Usage can be opaque
  - Modifications in progress
  - More user-friendly module flow
  - Based on interactions with key customers



DOCUMENTATION

HPE Cray Programming Environment

HPE Cray Programming Environment (CPE) suite offers programmers a comprehensive set of tools for developing, porting, debugging, and tuning applications. The programming environment simplifies the transition to new hardware architectures and configurations by automatically applying optimizations on HPC applications that use existing programming models with a simple recompile.

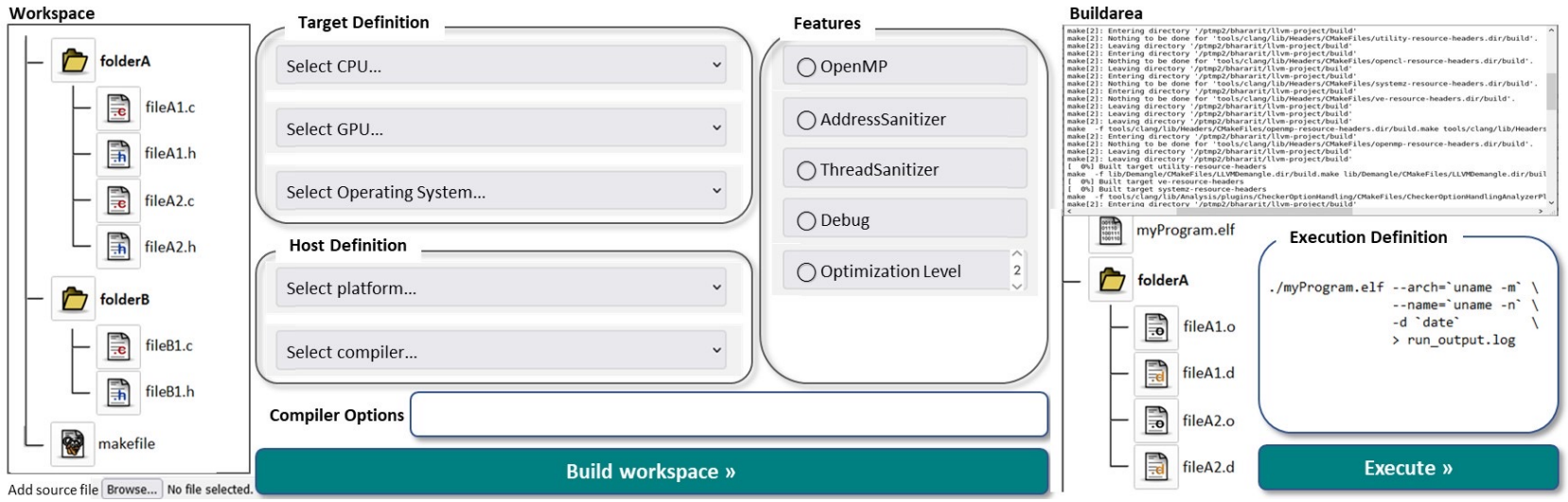
The environment provides:

- User environment (compiler drivers, hugepages, craype-api)
- Compilers, programming languages, and models
- Scalable communication libraries:
  - MPICH
  - PMI
  - OpenSHMEMX
  - DSMML
- Scientific and math libraries
- Debugging tools

TRAINING AND TUTORIALS

- CPE for scalable, portable application development
  - Introduction to development environment
  - Virtualized interface
  - Build systems managed in the backend
- Simplified selection of build tools
  - Compiler & version, platform
  - Target platform, capabilities, etc.
- Simplified build parameters
  - Debug, optimization, language features, sanitizers, etc.

- Selection of build hosts, tools & target platforms
  - Experiential learning for academics
  - Evaluation of migration & portability across platforms and compilers
  - Seamless triage & compiler capability verification



FUTURE OF COMPUTING: DIVERSITY AND INCLUSION

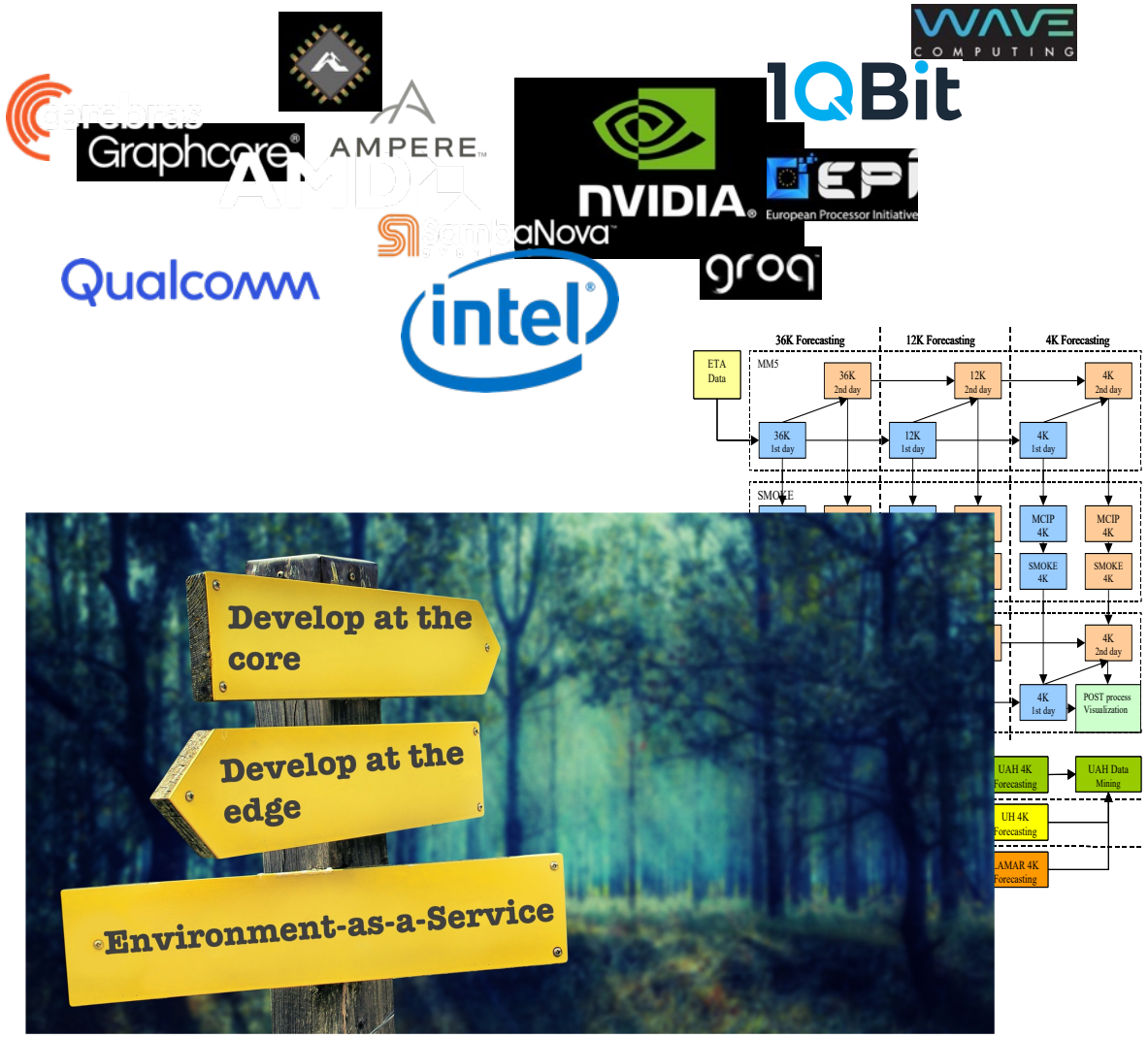
Applications?
- Workflows!
- Traditional simulation combined with AI

Platforms
- More integration on the node; customization via chiplets
- Specialized AI, Quantum devices integrated into HPC platform

Edge-to Exascale
- Development on laptops, on-prem systems, in Cloud
- Deployment across disparate resources

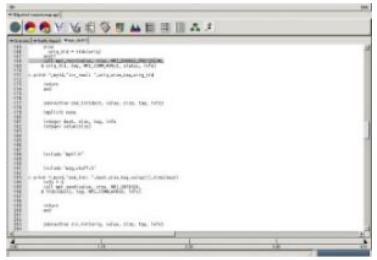
Programming languages and models
- Python, Julia, SYCL; DSLs? New languages?
- How will LLM transform the way we develop code??

Efficient execution
- Managing power and energy usage
- CPE aaS; dynamic program adaptation



HPE/CLASSIQ HYBRID HPC/QUANTUM COMPUTATION FLOW

Classiq

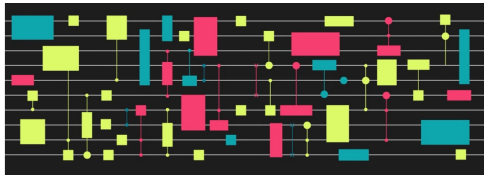


User Code +
CSML Library

CCE Compiler

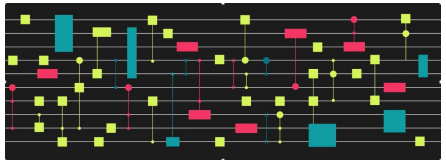
Cray Programming Env

Classiq
Quantum
interface

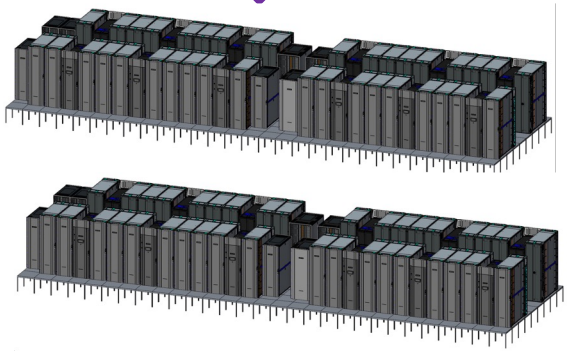


Classiq Synthesis
Engine

Output



Quantum Circuit

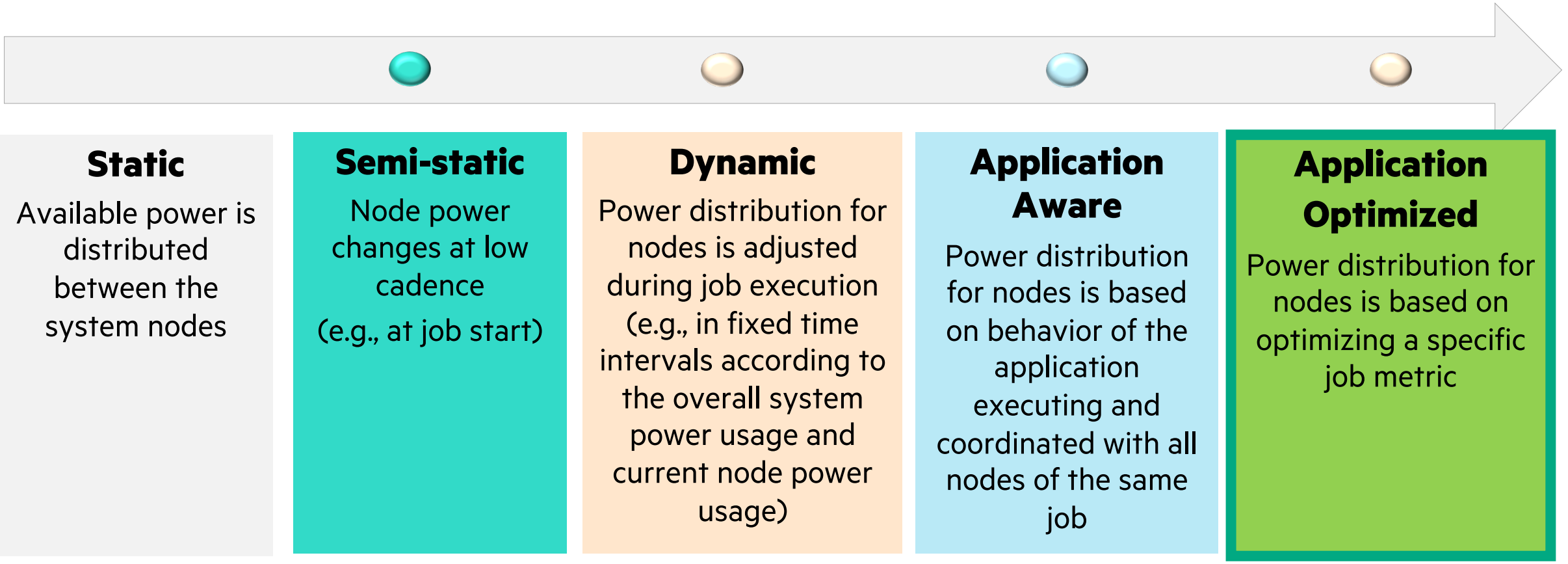


Quantum Simulators
Cray MPI

Simulator Output

Job launched using HPE's Workload Manager

PROGRESSION OF POWER MANAGEMENT FUNCTIONALITY



THANK YOU

Barbara.Chapman@hpe.com