STEP

SUSTAINABLE TOOLS ECOSYSTEM PROJECT

# Primer for Breakout #1: The Exploding Hardware Challenge

8/16/2023

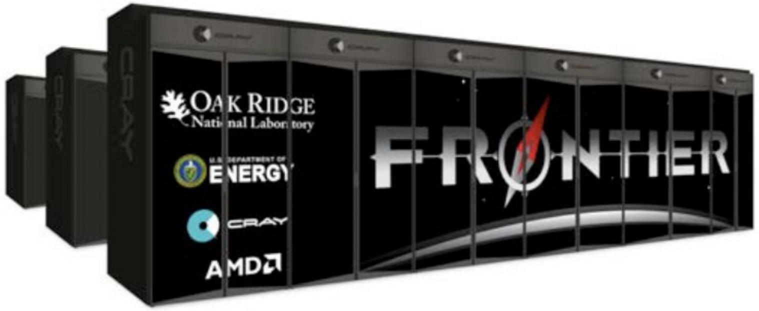
Michael Jantz, University of Tennessee

# Outline

- Exploding Hardware Complexity (John Mellor Crummey's talk)
  - Architectural diversity in emerging platforms
  - Increasing complexity in system design
  - Implications for Tools
- Breakout Sessions
  - Support Obstacles
  - Hardware Coverage
  - Vendor Engagement
  - Event Correlation
- Tasks for today

# Architectural Diversity in Emerging Platforms

- DOE's Emerging GPU-Accelerated Exascale Platforms
  - Frontier (OLCF) -- 9,472 compute nodes
    - 1 AMD EPYC "Trento" CPU
    - 4 MI250X AMD Radeon Instinct GPUs
    - 4 Slingshot 11 endpoints
    - Unified memory architecture
  - Aurora (ALCF) – 10,624 compute nodes
    - 2 Intel Xeon "Sapphire Rapids" processors
    - 6 Intel "Ponte Vecchio" GPUs
    - 8 Slingshot 11 endpoints
    - Unified memory architecture
  - El Capitan (LLNL) compute nodes
    - AMD MI300 APU: CDNA3 GPUs, Zen 4 CPUs, cache and HBM chiplets
    - Slingshot 11



Argonne National Laboratory
U.S. Department of Energy
intel
Hewlett Packard Enterprise
Aurora

U.S. Department of Energy
NNSA
Lawrence Livermore National Laboratory
AMD
Cray
EL CAPITAN

STEP SUSTAINABLE TOOLS ECOSYSTEM PROJECT

Slide adapted from STEP: East Coast Town Hall presentation by Mellor-Crummey [1]

# Architectural Diversity in Emerging Platforms

- Other Emerging Platforms
  - Cerebras Andromeda (11/2022)
    - >1 ExaFLOP of AI compute, >120 PetaFLOPs of dense compute
    - 16 Cerebras CS-2 systems (13.5 million AI-optimized cores)
    - 8,176 AMD Epyc Gen 3 cores
    - 96.8 terabits-per-second node-node fabric bandwidth

  - NVIDIA Helios (announced 5/2023)
    - NVIDIA DGX GH200: 256 Grace Hopper PEs connected by 2-level NVLINK network

    - > 1 exaflops of FP8 AI performance (or ~9PF of FP64 performance)
    - Helios: 4 DGX GH200 connected by NVIDIA Quantum-2 Infiniband

STEP SUSTAINABLE TOOLS ECOSYSTEM PROJECT

Slide adapted from STEP: East Coast Town Hall presentation by Mellor-Crummey [1]

# Increasing Complexity in System Design

- CPU-based architectures
  - Heterogeneous cores
    - Energy efficient / high performance cores
  - Multi-level memories
    - Deeper cache hierarchies
    - On-package / off-package memory
    - Low-power, non-volatile memories
  - On-chip accelerators
    - SIMD computation
    - Data movement
    - Compression, cryptography, etc.

Slide adapted from STEP: East Coast Town Hall presentation by Mellor-Crummey [1]

# Increasing Complexity in System Design

- GPUs
  - Several manufacturers now offer powerful discrete GPUs
    - NVIDIA GPUs: SIMT organization
    - AMD GPUs: scalar + SIMD vector operations together
    - Intel GPUs: SIMD operations, with scalar operations in a single SIMD lane
    - Memory hierarchies in each architecture are also different

Slide adapted from STEP: East Coast Town Hall presentation by Mellor-Crummey [1]

# Increasing Complexity in System Design

- Integrated architectures
  - AMD MI300 APU
    - CDNA3 GPUs, Zen 4 CPUs, cache and HBM chiplets

- Special purpose architectures
  - Mainly for AI / ML training
    - Habana Gaudi2
    - Cerebras CS-2
    - Sambanova Datascale
    - Groq GroqChip

Slide adapted from STEP: East Coast Town Hall presentation by Mellor-Crummey [1]

# Increasing Complexity in System Design

- Interconnects
  - Node-level
    - PCIe-5, NVIDIA NVLINK, AMD Infinity Fabric, Intel Ultra Path Interconnect
    - May use explicit copies (e.g., memcpy) or implicit copies (e.g., page faults) to move and communicate data within a node
  - System-level
    - NVIDIA NVLINK, HPE/Cray Slingshot, Mellaknox Infiniband
    - Use messaging protocols (e.g., MPI, RDMA) to communicate data between nodes

Slide adapted from STEP: East Coast Town Hall presentation by Mellor-Crummey [1]

# Increasing Complexity in System Design

- File Systems and I/O
  - Storage system technologies
    - Parallel file system target nodes with disks
    - Solid state burst buffer (e.g., Frontier)
    - Solid state data store (e.g., Optane on Aurora, flash on Perlmutter)
  - File system abstractions
    - Lustre
    - Distributed Asynchronous Object Storage (DAOS)
      - open-source object store for massively distributed non-volatile memory
    - IBM Spectrum Scale, formerly GPFS

Slide adapted from STEP: East Coast Town Hall presentation by Mellor-Crummey [1]

# Implications for Tools

- Tools are needed to identify:
  - where an application spends its time (bottlenecks)
  - what resources are available
  - how well the system and applications are using these resources
  - opportunities for improvement
- Explosion of architectural variety presents daunting challenges for tools
  - Profusion of HW counters
  - Incomplete capabilities for measurement and attribution
  - Architecture dependent strategies for analyzing and diagnosing

STEP
SUSTAINABLE TOOLS
ECOSYSTEM PROJECT

Slide adapted from STEP: East Coast Town Hall presentation by Mellor-Crummey [1]

# Breakout Sessions on Exploding HW Complexity

- Breakout #1: Support Obstacles
- Breakout #2: Hardware Coverage
- Breakout #3: Vendor Engagement
- Breakout #4: Event Correlation

# Breakout #1: Support Obstacles

- Objective: 1) identify the most significant obstacles to supporting new hardware or platforms, and 2) identify mechanisms that are most helpful for addressing these obstacles
- Biggest obstacles to supporting new hardware or platforms?
  - Developers only have hands-on access to a machine after it has already been accepted at a facility ("reactive cycle" problem)
  - Tension between wanting to provide *general availability* of a machine as quickly as possible and need for coordination between key personnel at facilities and vendors
  - Lack of coordinated forum for communication among stakeholders
  - Challenges with familiar human traits (e.g., denial, resistance to change)

# Breakout #1: Support Obstacles

- Mechanisms to address these obstacles?
  - Three possible mechanisms identified:
    - Vendor □ community briefings
    - Early and ongoing test access
    - Open communication channels
  - Several existing methods also noted, including:
    - Training for end users
    - Hackathons

# Breakout #2: Hardware Coverage

- Objective: identify challenges and possible solutions for how to achieve tools sustainability in terms of hardware coverage
- Hardware coverage gap concerns(according to time frame)
  - Immediate term: GPU tool support
  - Near term (< 5 years): tool support for accelerators, FPGAs throughout the system architecture and along the data path
- Other cross-cutting hardware coverage gaps:
  - Tools for power management
  - Tools for load balancing across heterogeneous capabilities
- User concerns: how to ensure users find the “right tool for the right job”

# Breakout #2: Hardware Coverage

- Recommendations for achieving tool sustainability in the face of increasing hardware complexity:
  - Contingency funding (e.g., to help the community react more quickly to new HW)
  - Standardization of counters (impractical broadly, but would be useful if targeted)
  - Considering alternatives to counters for HW instrumentation
  - Availability of "mini-apps for tooling"
  - Early access to hardware
  - Open source hardware designs
  - FOA mandates that support HPC tools
  - Better community coordination to ensure coverage (top down documentation, sharing profile data)

# Breakout #3: Vendor Engagement

- Objective: explore how vendors can contribute to tool sustainability and determine how tools can support vendors in this initiative
- Key recommendations:
  - Vendors should proactively disclose forthcoming vendor-specific software and hardware to tool developers. Potential benefits include:
    - External feedback from tools can help vendors enhance their own products
    - Relieves vendor of burdens of tool development
    - Timely availability of tools on new machines
  - Tool developers should engage in proactive communication with vendors prior to releasing vendor-specific support within their toolsets
  - Establish a "Vendor-Tools Alliance" workshop and website
    - Workshop to facilitate communication between vendors and tool developers
    - Website could serve as a hub for collecting and consolidating issues / feedback from various tools

# Breakout #4: Event Correlation

- Objectives:
  - 1) identify key use cases for correlating hardware events with program execution and/or application source code
  - 2) identify technologies and capabilities that are needed to improve correlation between applications and hardware and ensure it can be achieved sustainably
- Key use cases for event correlation
  - Correctness tools need to attribute problems to program context (including source code, program variables)
  - Debugging tools need to correlate program behavior with code positions and values of program variables
  - Performance tools need to be able to correlate program behavior with source code context, program data, and hardware resources to be able to assess the impact of application behavior on resource utilization

# Breakout #4: Event Correlation

- Key gaps in event correlation
  - New architectures often lack necessary HW and SW mechanisms for monitoring hardware events and correlating them with source code positions
  - New architectures often lack documentation, such as top-down models, that inform how to use the available HW mechanisms for measurement and event correlation
  - GPU architectures have under-developed correlation mechanisms (e.g., GPUs cannot attribute stalls to their root causes, or identify the level of the memory hierarchy from which a load is satisfied)
  - Mechanisms to correlate network performance with network and I/O operations are also lacking

# Breakout #4: Event Correlation

- Key recommendations to improve event correlation
  - Subject matter experts should work independently from the procurement process to identify best practices and gaps in event correlation and publish these in a white paper (or platform capability standard)
    - Vendors will then be able to work to address these needs before offering them in a procurement
  - The HPC community should join forces with other relevant stakeholders (including cloud providers and other large data service providers) to:
    - Identify common needs and communicate these needs to technology developers
    - Identify gaps where new methods for better event correlation are needed

# Today's Breakout

- Rank key issues / tasks / recommendations based on:
  - Importance for sustainability
  - Likelihood of success
- Goals are to:
  - Summarize material covered in previous town halls
  - Refine and prioritize the set of tasks / recommendations so that we can organize them into a plan for our proposal

# References

[1] Mellor-Crummey, John. Tools Challenge: Exploding Hardware Complexity *CPUs, GPUs, and Accelerators, Oh my!* Presentation at STEP East Coast Town Hall. New York, NY. June, 2023.