



STEP

SOFTWARE TOOLS
ECOSYSTEM PROJECT

Introducing STEP: A Presentation to the Facilities Software Task Force

January 8, 2024

Terry Jones, STEP Director

Introducing the Software Tools Ecosystem Project (STEP)

- STEP In A Nutshell
- The Tools Ecosystem and Why It's Special
- The 5 Initial Tools under STEP stewardship
- A look toward next steps



STEP in a nutshell...

Lead PI: Terry Jones
Deputy PI: Phil Carns



Objectives and Scope

- Tools and supporting software for monitoring, analysis, and diagnosis of performance and behavior of codes on advanced computing systems.
- Examples: application profilers, tracing tools, system monitors, etc.
- Co-design with hardware vendors, application developers, facilities and tool developers.

Challenges

- These tools rely on a deep understanding of hardware, and hardware is rapidly changing.
- Vendors are reluctant to divulge hardware details.
- Stakeholders are very diverse and currently have insufficient communication paths.

Initial Funded Software Packages

Software Tool	PI
HPCToolkit	John Mellor-Crummey, Rice Univ.
PAPI	Heike Jagode, Univ. of Tennessee
Dyninst	Barton Miller, Univ. of Wisconsin
Tau	Sameer Shende, Univ. of Oregon
Darshan	Shane Snyder, Argonne

Opportunities

- STEP can be a vehicle to improve communication among tool developers, application teams, vendors and facilities.
- STEP can be a clearing-house for efficiently managing critical HPC technology.
- STEP can be a resource for workforce development.



STEP Specifically Targets *Tools*

WHAT ARE TOOLS?

- We define Tools to mean *“the collection of tools and utilities that can be applied to both understanding performance bottlenecks and facilitating run time mitigation of performance degrading phenomena.”*
- These tools interact with hardware, compilers, communication libraries, programming model runtimes, operating systems, middleware layers, as well as the many applications that utilize these tools.

What's *So Important* about Tools?

- *As computers have increased in complexity and scale, using them effectively has become much more difficult.*
- In addition to their role in enabling supercomputer performance (a decisive determinant of scientific discovery), these tools provide essential feedback to users, operations staff, and system and application software developers.



How Are Tools **Unique**?

- Tools are closely bound to architectures and system software in ways that other types of software, such as libraries and scientific applications, are not.
- For example, a tool that tracks how an application uses computing resources must be able to measure low-level architectural events and metrics and relate them to program progress and source code.
- The need for tools is most acute for understanding code performance on systems that push the boundaries of technology and scale, but these systems' novelty makes them extremely difficult for tool developers to support when first deployed.

STEP and Facilities

- STEP was designed with Facilities in mind: STEP facility representatives participated in the development of STEP as investigators

Richard Gerber, LBNL	Kevin Harms, Argonne	Matthew Legendre, LANL
Susanne Parete-Koon, ORNL	Galen Shipman, LANL	Veronica Melesse Vergara, ORNL

- Supporting the activities that are named in the initial Y1 plans that are heavily geared towards sustainability – See key activities for HPCToolkit, PAPI, Dyninst, TAU and Darshan.
- A convenient interface when Facilities needs to broadly reach the tools community with a high-level concern.
- A mechanism to address needs and requirements: give STEP input on things that are important for the tools portfolio for out years, STEP will seek software tool solutions.
- A source for tools information and assistance (e.g., software aspects of machine procurement).

PAPI: Performance API

What is PAPI?

- PAPI provides a **universal interface** and methodology for **monitoring performance counters** and **managing power** across a diverse range of **hardware** and **software** components.

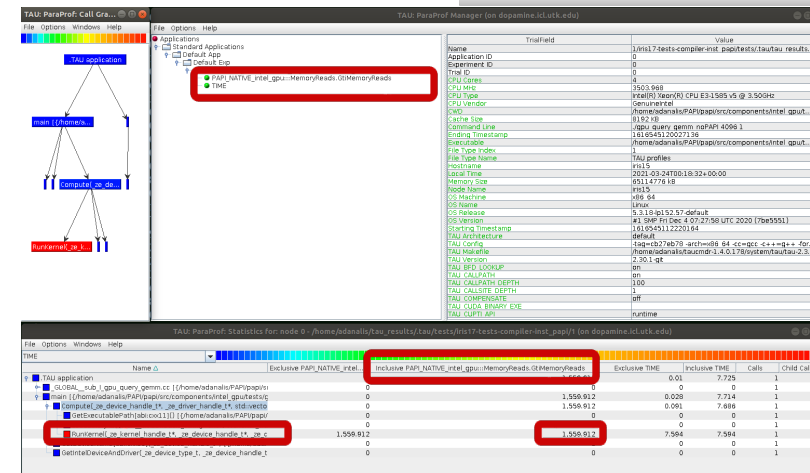
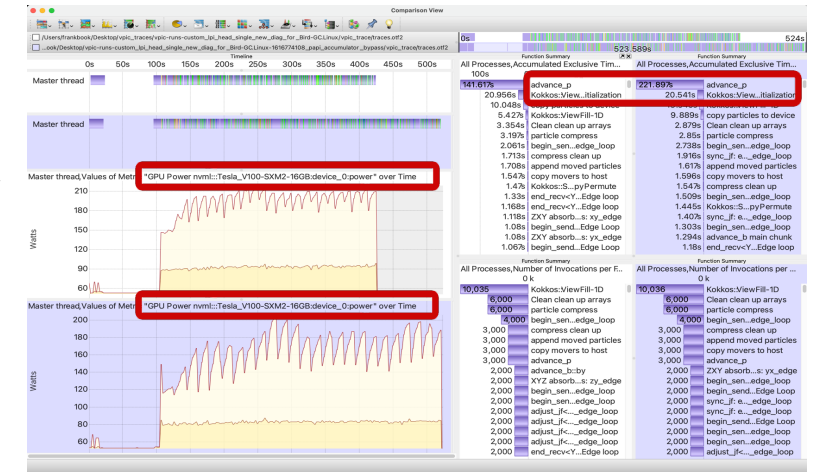
Monitoring Capabilities

- Robust support for **advanced hardware** resources, which includes major CPUs, GPUs, accelerators, interconnects, I/O systems, power interfaces, and virtual cloud environments.
- Finer-grain power management** for dynamic optimization of applications under power constraints.
- Software-Defined Events** originating from software stack (e.g., communication and math libraries, runtime systems).
- Semantic analysis** of hardware counters so that scientists can better make sense of the ever-growing list of events.

Community Impact

- PAPI channels the monitoring capabilities of **hardware counters**, **power usage**, and **software-defined events** into a robust software package for HPC systems.
- PAPI provides the ability to monitor these metrics through a **single, portable API**.
- Without PAPI, applications must use multiple APIs to access all available metrics, which can hinder productivity.

PAPI performance and power measurements of a Kokkos application using Vampir / Score-P on NVIDIA GPUs.



PAPI performance measurements of a math kernel using TAU on Intel GPUs.

Expected Future Impact from PAPI

- Monitoring capabilities of **new and advanced technologies** across the hardware-stack, e.g., GPU-to-GPU interconnect support for novel GPUs, rocprofiler v2 support for AMD GPUs.
- PAPI abstractions for expressing the **internal behavior of software** components and improved **interoperability** across the software-stack.



Slide source: Heike Jagode

<https://ascr-step.org>

presentation to the Facilities Software Task Force

DyninstAPI: Binary Code Analysis and Instrumentation Toolkit

Current Capabilities on CPU Architectures

Includes x86, ARM and Power

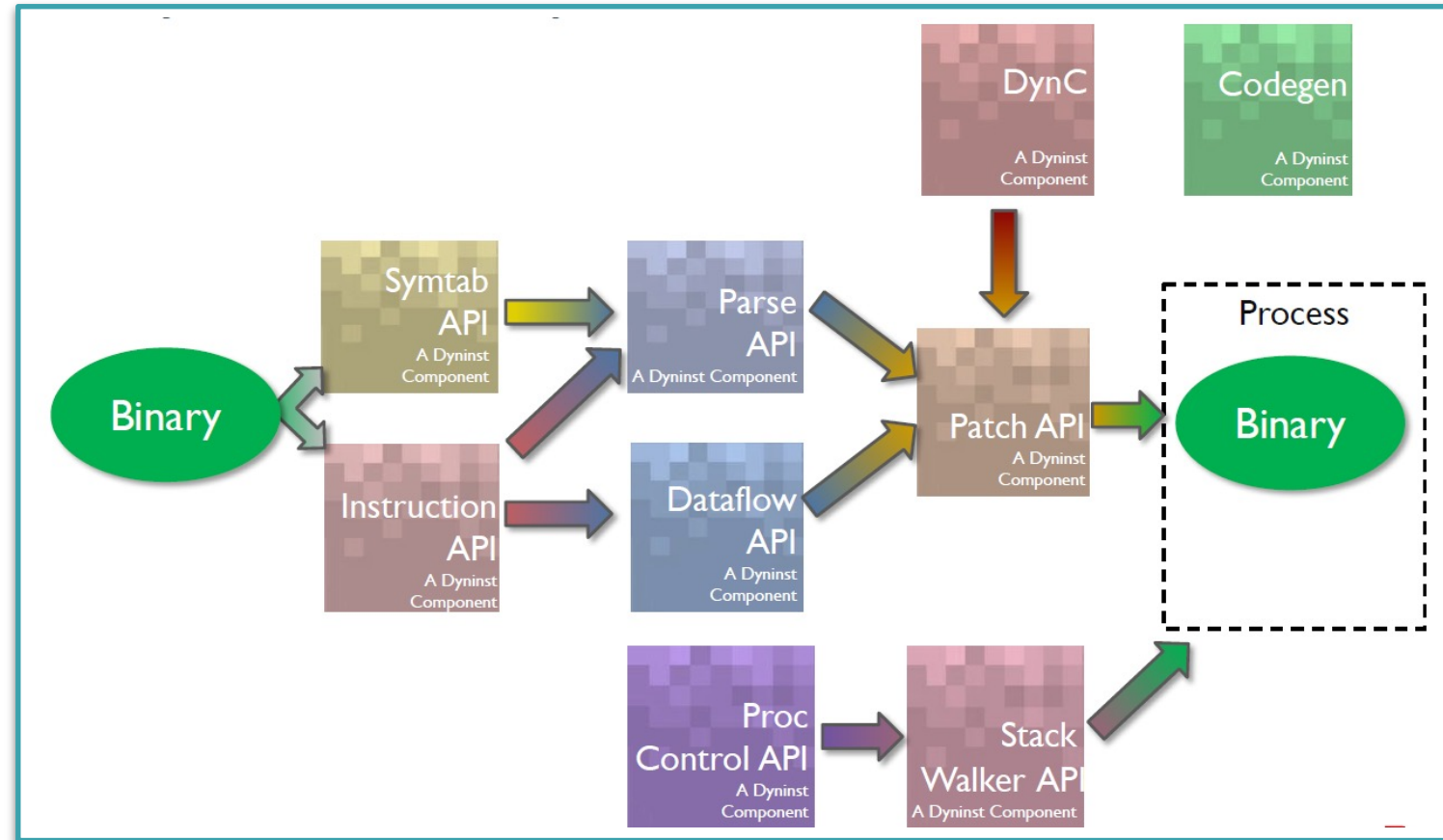
Binary Code Analysis

- Abstract program representation avoids the need for architecture specific knowledge
- Full control flow analysis of binary code to identify functions, loops, and basic blocks
- Dataflow analysis to support slicing, symbolic evaluation and pointer analysis
- Operates on stripped code

Binary Code Instrumentation

- Both static (modify executables and libraries) and dynamic (modify running applications)
- Supports dynamically and statically linked code
- High level: Can instrument function entry, exit and call; loop entry, exit and top;
- Portable: DyninstAPI tools can work across any of the supported architectures.
- Fine grained: Can instrument at any instruction
- Efficient

Is the instrumentation foundation for HPCToolkit, AMD Omnitrace, TAU, Cray ATP, Red Hat SystemTap, and others



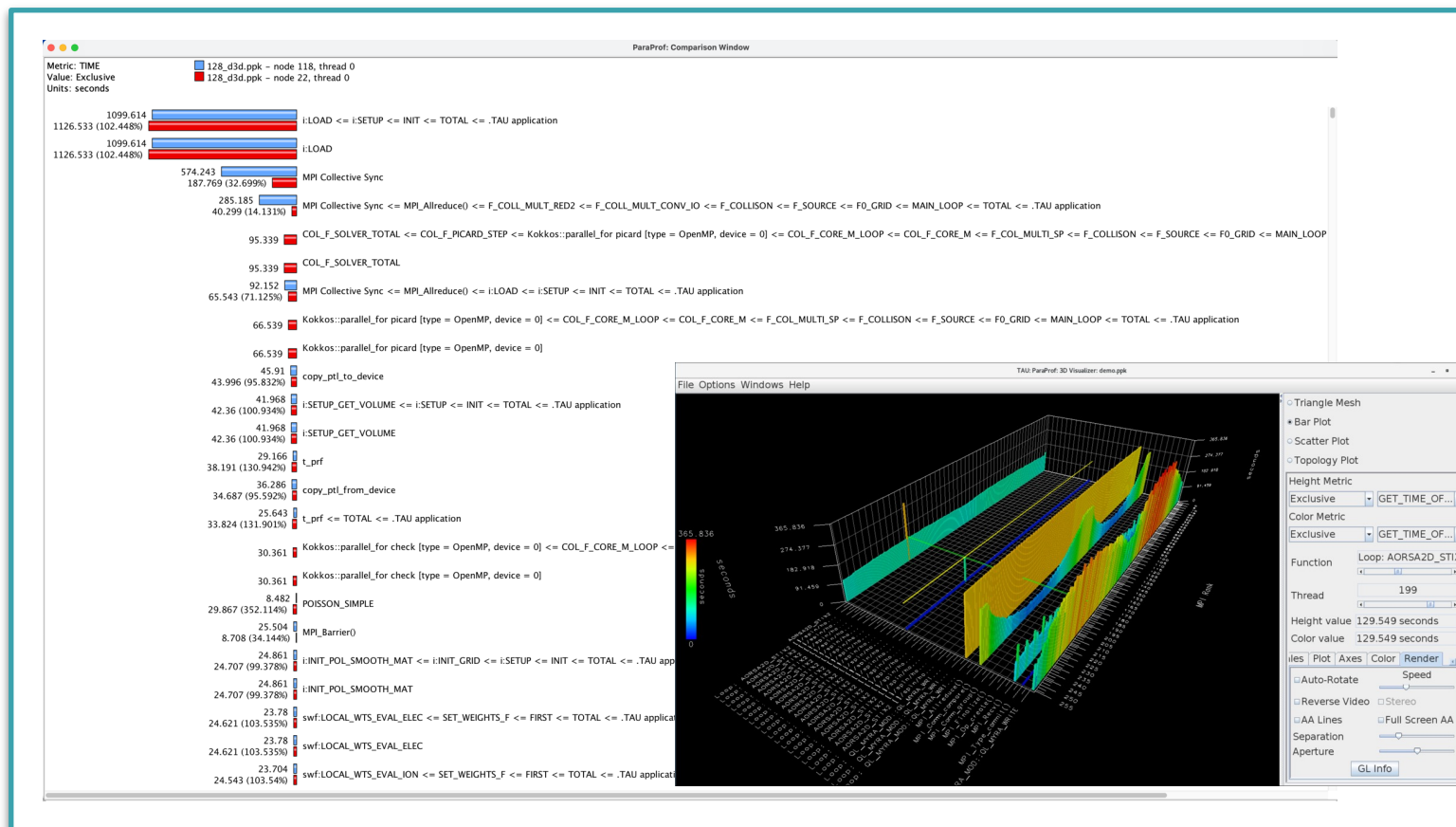
Key Plans for 2024

- Full Dyninst support for the newest x86 CPU families
- Extend Dyninst to support binary instrumentation of AMD GPU

TAU Performance System®: Performance Evaluation Tool for GPU-accelerated Architectures

Current Capabilities

- Integrated performance instrumentation, measurement, and analysis toolkit
- TAU is installed as a module at ALCF, LLNL, OLCF, LANL, Sandia, NERSC, and other supercomputing centers
- Supports DyninstAPI for instrumentation, PAPI for hardware performance counters based measurement, HPCToolkit and Darshan for profile display in TAU's 3D profile browser, ParaProf
- TAU supports Python instrumentation (TensorFlow, PyTorch, Horovod) and HPC runtimes (Kokkos, MPI, OpenMP)
- TAU supports tracing with OTF2 and visualization using Vampir (TU Dresden)
- `tau_exec` launcher requires no modification to the application binary!
- On AMD GPUs, TAU supports OMPT, ROCProfiler, ROCTracer, LLVM plugin for compiler-based instrumentation using hipcc, ROCTX
- On NVIDIA GPUs, TAU supports CUDA Profiling Tools Interface (CUPTI), NVTx
- On Intel GPUs, TAU supports OMPT, Intel Level Zero and OpenCL, Kokkos



Key Plans for 2024

- *Expand TAU's support for rewriting binaries and Dynamic Shared Objects using DyninstAPI*
- *Expand CI/CD support for TAU using GPUs from Intel, AMD, and NVIDIA on Frank at UO*



Darshan: Enabling Insights into I/O Behavior of HPC Applications

Darshan is an application I/O characterization tool commonly deployed by HPC facilities

- Lightweight instrumentation methods to avoid perturbing applications
- Transparently enabled, requiring no modifications to users' code
- Generates summary reports of applications' I/O behavior

1 Darshan runtime library

Transparently intercept application I/O calls, capture I/O characterization data, and generate log file when application terminates

Existing instrumentation modules span the HPC I/O stack: HDF5, PnetCDF, MPI-IO, POSIX, Lustre, etc.

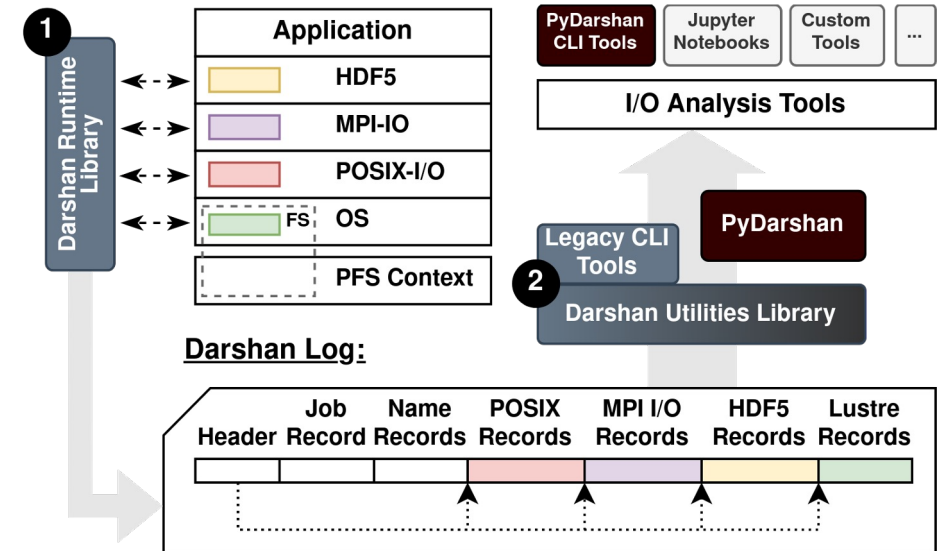
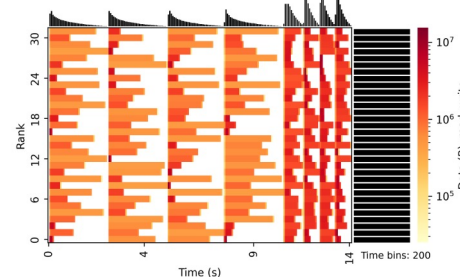
2 Darshan log analysis tools

New PyDarshan framework to enable extraction and analysis of Darshan log data using popular Python packages like pandas, matplotlib, etc

PyDarshan includes a job summary report tool to provide I/O overviews for users' logs

- E.g., I/O performance estimates (left) and I/O activity heatmaps (right)

files accessed	1026
bytes read	50.10 GiB
bytes written	49.30 GiB
I/O performance estimate	164.99 MiB/s (average)



Key Plans for 2024

- Extend Lustre FS instrumentation to account for new progressive file layout (PFL) feature
- Improve design and capabilities of PyDarshan analysis tools
- Enhance CI testing capabilities using facility resources

A Look Forward

Can Tools Transition from Reactive to Proactive?



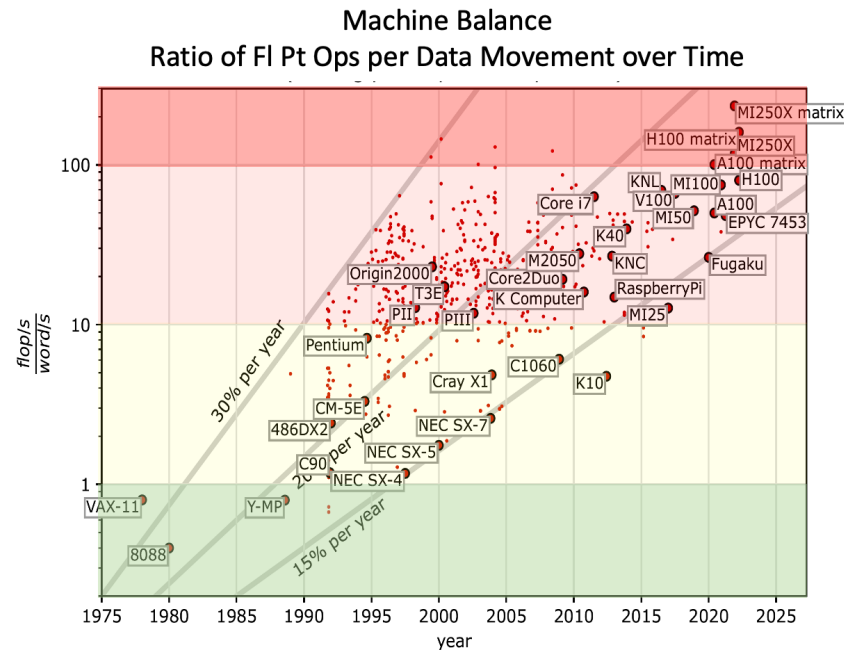
A Look Forward

Can Tools Transition from Reactive to Proactive?

Can Tools Address the Memory Imbalance?

When We Look at Performance in Numerical Computations ...

- Data movement has a big impact
- Performance comes from balancing floating point execution (**Flops/sec**) with memory->CPU transfer rate (**Words/sec**)
 - “Best” balance would be 1 flop per word-transferred
- Today’s systems are close to 100 flops/sec per word-transferred
 - Imbalanced: Over provisioned for Flops



Plot for 64-bit floating point data movement & operations
(Bandwidth from CPU or GPU memory to registers)

*Source: Jack Dongarra, <https://new.nsf.gov/events/overview-high-performance-computing-future>

A Look Forward

Can Tools Transition from Reactive to Proactive?



Can Tools Address the Memory Imbalance?

When We
Numerical

- Data movement
- Performance compared to floating point execution with memory -> C (Words/sec)
 - “Best” benchmark: 1 flop per word
- Today’s systems: 100 flops/sec per word
 - Imbalance
- Over provisioned

What’s The Role for Tools In The Long Term? (Matt Welsh @ Harvard)

Is Computer Science Doomed?

Computer science
1959-2030

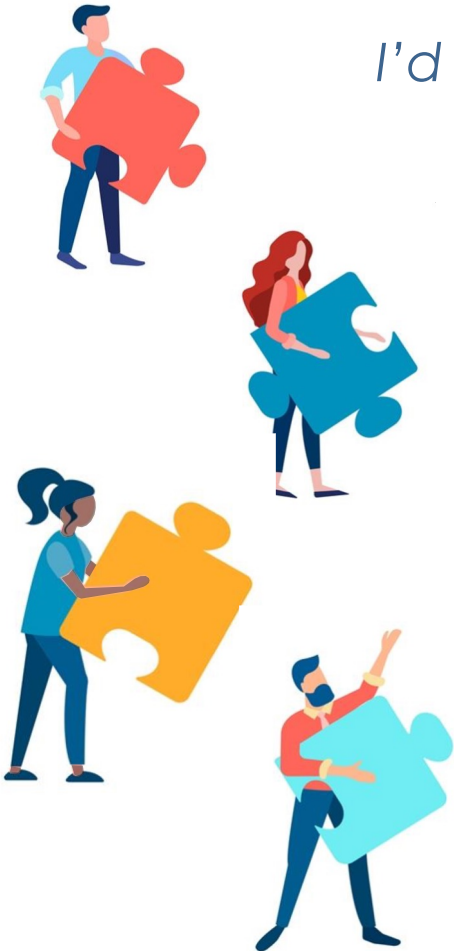
Source: The New Stack,
<https://thenewstack.io/if-computer-science-is-doomed-what-comes-next/>

QUESTIONS?



I'd Like to acknowledge the full Software Tools Ecosystem Project (STEP) Team

NAME	PERSPECTIVE	NAME	ORGANIZATION
Lead PI	Tools	Terry Jones	Oak Ridge National Lab
Deputy PI	Tools	Philip Carns	Argonne National Lab
Co-PI	Tools	James Brandt	Sandia National Laboratories
Co-PI	Vendors	James Custer	Hewlett Packard Enterprise
Co-PI	Tools	Ann Gentile	Sandia National Laboratories
Co-PI	Facilities	Richard Gerber	Lawrence Berkeley National Lab
Co-PI	Facilities	Kevin Harms	Argonne National Lab
Co-PI	Tools	Heike Jagode	University of Tennessee
Co-PI	Tools	Mike Jantz	University of Tennessee
Co-PI	Tools	Matthew Legendre	Lawrence Livermore Natl Lab
Co-PI	Vendors	John Linford	NVIDIA
Co-PI	Vendors	Keith Lowery	Advanced Micro Devices
Co-PI	Facilities	Verónica Melesse Vergara	Oak Ridge National Lab
Co-PI	Tools	John Mellor-Crummey	Rice University
Co-PI	Tools	Barton Miller	University of Wisconsin
Co-PI	Vendors	José Moreira	IBM
Co-PI	Applications	Erdal Mutlu	Pacific Northwest Natl Lab
Co-PI	Facilities	Suzanne Parete-Koon	Oak Ridge National Lab
Co-PI	Tools	Sameer Shende	University Oregon
Co-PI	Tools	Shane Snyder	Argonne National Lab
Co-PI	Tools	Galen Shipman	Los Alamos National Laboratory
Co-PI	Tools	Devesh Tiwari	Northeastern University
Co-PI	Applications	Theresa Windus	Ames National Lab



Extra Slides

STEP Coverage Matrix – Software Packages Fostered Through STEP

	Current Machines	Future Machines (STEP)
Meta Tool used by other tools	4, 6, 10	4, 6, 10
CPU Examination & Performance	2, 7, 12	7
GPU Examination & Performance	2, 7, 12	7
Memory Examination & Perf	9, 11	
Storage I/O Examination & Perf	4, 5, 8	4
Debugger & Correctness	13	
Parallel load-balance / scaling	7, 12	7
Machine Level Exam & Perf	1, 4, 7, 12	7, 12
Network/Workflow level Exam & Perf	3	

Legend (green=funded):

- 1=ArgoNRM,
- 2=Bliss,
- 3=Chimbuko,
- 4=Darshan,
- 5=Drishti,
- 6=Dyninst,
- 7=HPCToolKit,
- 8=LDMS Conn,
- 9=Memgaze,
- 10=PAPI,
- 11=SICM,
- 12=TAU,
- 13=Thapi

CY2024 STEP Portfolio

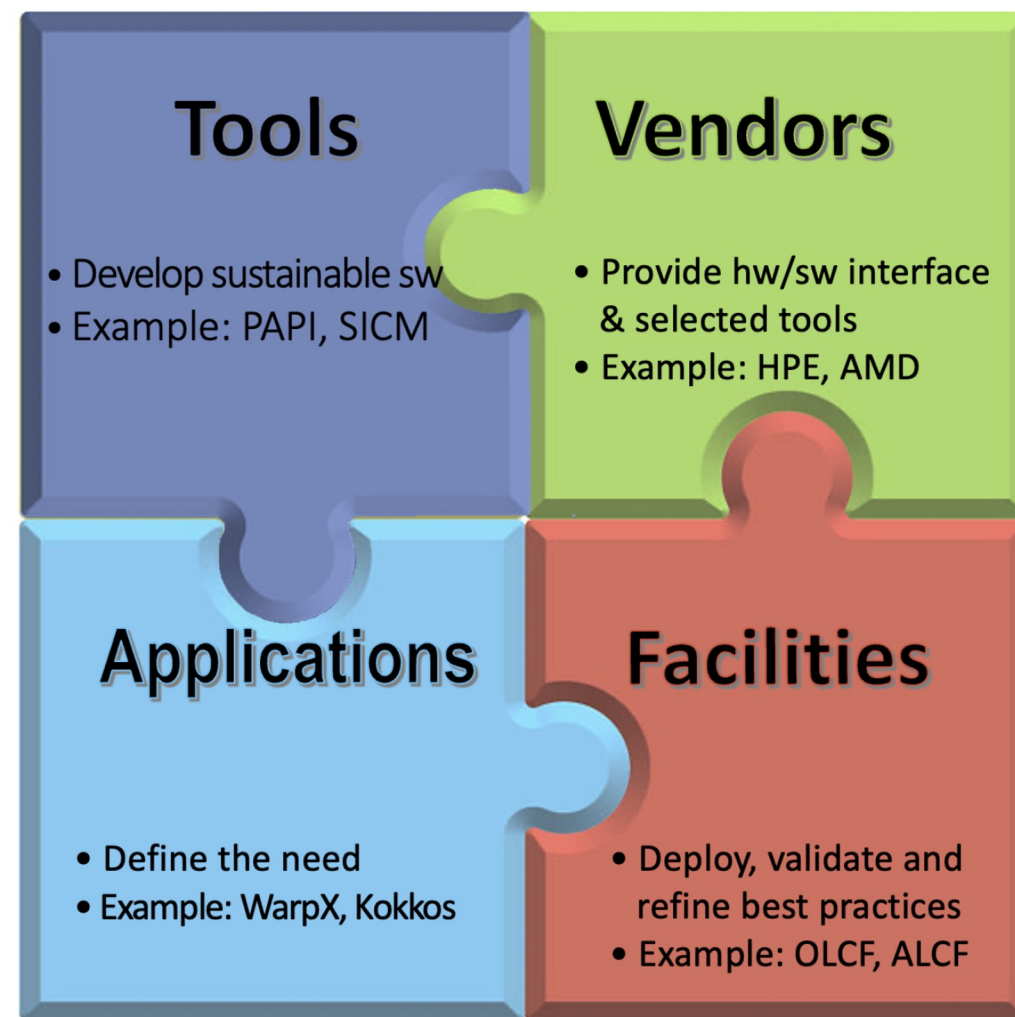
Tool	Tool Description	Tool PI	Lead Institution	STEP Funding	Former ECP Funded	DOE Deployed	Maturity ¹
HPCToolkit	Performance analysis toolkit for HPC	John Mellor-Crummey	Rice University	Yes	Yes	Yes	M
PAPI	Perf monitoring library & toolset	Heike Jagode	Univ. of Tennessee	Yes	Yes	Yes	M
Dynist	Binary code and instrumentation	Barton Miller	Univ. of Wisconsin	Yes	Yes	Yes	M
TAU	Portable profiling and tracing toolkit	Sameer Shende	Univ. of Oregon	Yes	Yes	Yes	M
Darshan	I/O instrumentation tool	Shane Snyder	Argonne Natl Lab	Yes	Yes	Yes	M
Additional critical software tools that STEP would support with a larger budget ceiling.							
SICM	Memory tool for multi-tier memories	Terry Jones	Oak Ridge Natl Lab	tbd	Yes	Yes	M
Drishfi	I/O Analysis and visualization tool	Suren Byna	Ohio State Univ.	tbd	Yes	Yes	M
Chimbuko	Tool for extreme scale machines	Christopher Kelly	Brookhaven Natl Lab	tbd	Yes	Yes	I
LDMSinspect	Darshan/LDMS integration	Devesh Tiwari	Northeastern Univ	tbd		Yes	I
Argo NRM	Node resource management tool	Swann Perarnau	Argonne Natl Lab	tbd	Yes	Yes	I
Thapi	Heterogeneous API instrumentation	Brice Videau	Argonne Natl Lab	tbd	Yes	Yes	I
Memgaze	Memory analysis toolset	Nathan Tallent	Pacific NW Natl Lab	tbd			I
Qraff	Quantum toolkit	Devesh Tiwari	Northeastern Univ	tbd			I

¹STEP intends to support both mature software products and newer emerging tools. Maturity is denoted by either M=mature or I=Incubator.



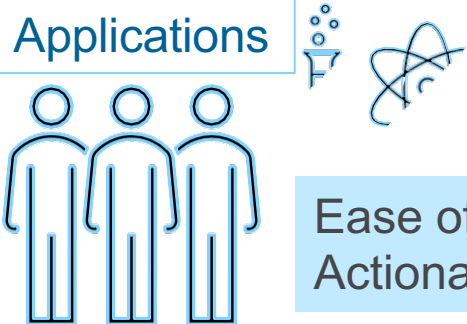
STEP Aspirations

- Support Tools Community
- Utilize Co-design
- A proactive stance on the rapidly evolving hardware landscape: new (completely different) tool APIs from AMD and Intel that include support for GPU PC sampling.




Tools Stakeholders And Their Concerns

Applications



Ease of use,
Actionable feedback

Vendors



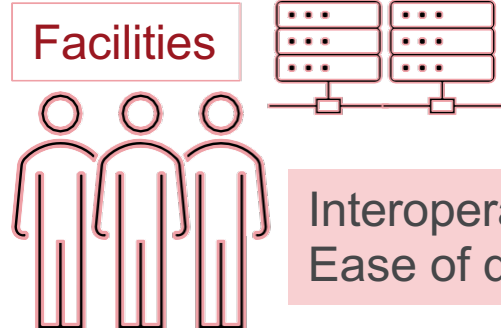
Maximize impact,
Product success

Tool Devs



Novel insights,
Broad applicability

Facilities



Interoperability,
Ease of deployment

Several significant attributes...



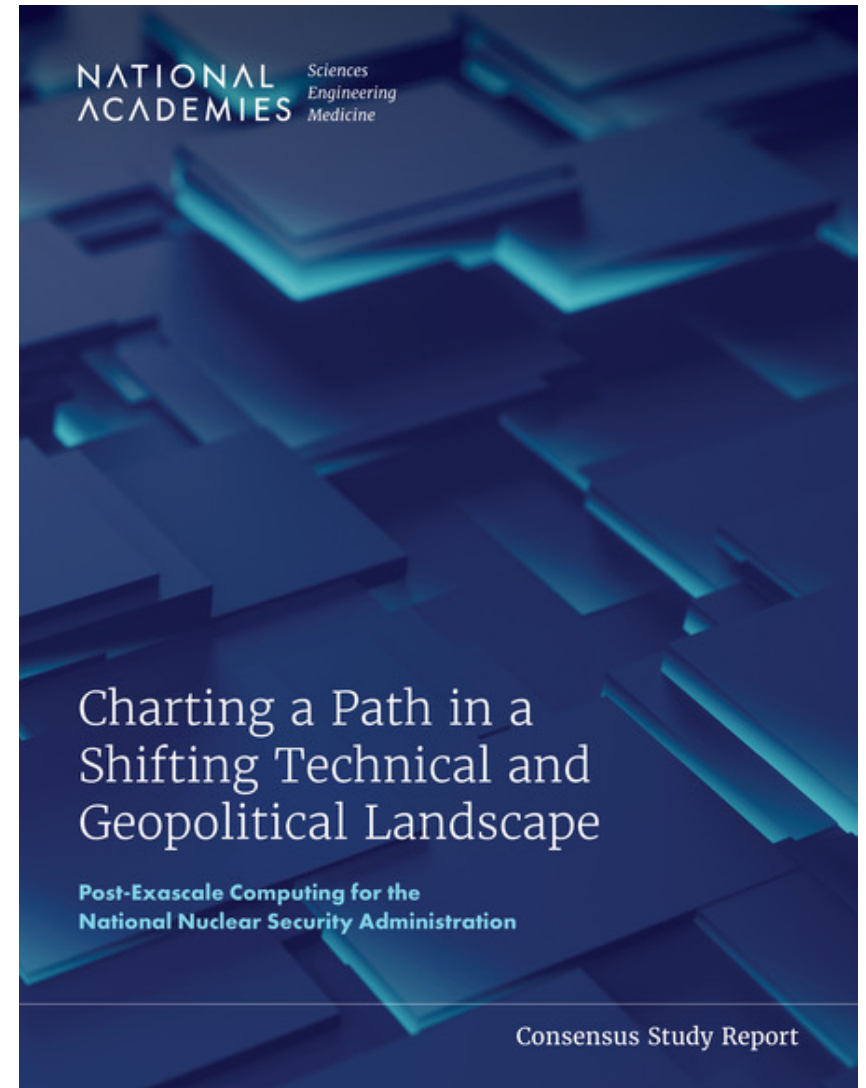
1. Not **_just_** replacing ECP
2. Focus on **_stewardship_**
3. Requires **_strong partnership_** with facilities, vendors, and application teams.

National Academies on Post-Exascale for NNSA...

“On the technology front, the single-thread performance of microprocessors continues to be relatively flat, and improvements in transistor density are slowing. Processing elements increasingly rely on more abundant, finer-grained parallelism and increasingly specialized hardware features that can improve performance by tailoring to a given computational domain. These trends are creating a significant disruption in available hardware components, with commercial interests focused largely on AI, embedded systems, and cloud services. **There is considerable uncertainty as to whether the processors emerging in response to these trends, with their lower precision and limited high-speed memory, can be productively applied to NNSA applications.**” (page 4 [emphasis mine])

“**FINDING 3.6:** Co-design of hardware and systems for high-performance scientific computing applications has been a modest success to date and will be more important in the future and need to be deeper. Technological and market trends are likely to shift the balance of co-design to the laboratories, requiring more innovation and engineering in the areas of hardware design, system integration, and system software.” (page 7)

“**RECOMMENDATION 2.2:** NNSA should strengthen efforts in computer science research and development to build a substantial, sustained, and broad-based intramural research program that is positioned to address the technological challenges associated with post-exascale systems and co-design of those systems to ensure that the laboratories are positioned for leadership in computing breakthroughs relevant to NNSA mission problems.” (page 8)



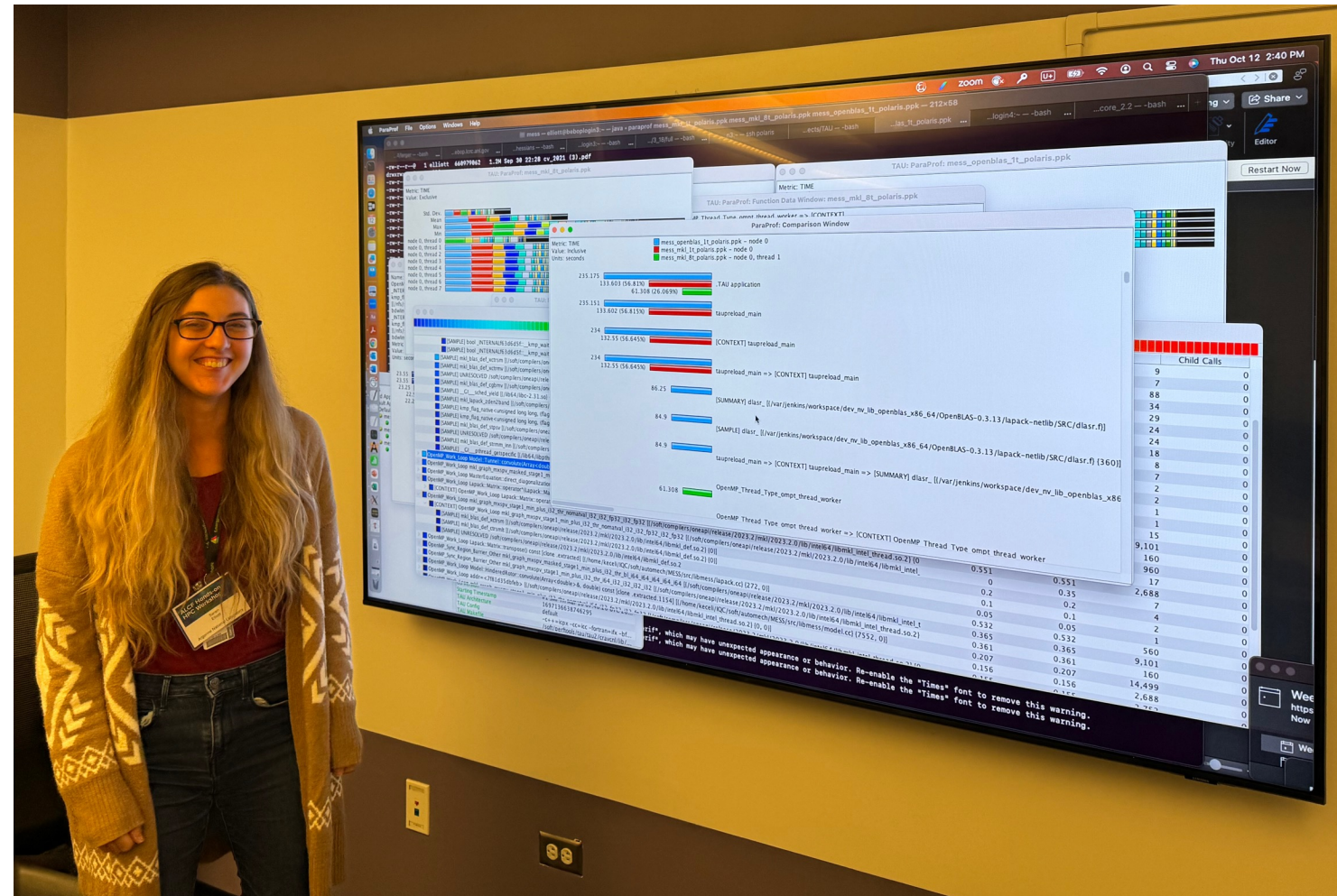
Source: National Academies of Sciences, Engineering, and Medicine. 2023. Charting a Path in a Shifting Technical and Geopolitical Landscape: Post-Exascale Computing for the National Nuclear Security Administration. Washington, DC: The National Academies Press. <https://doi.org/10.17226/26916>.



Tools Impact: A *Recent TAU Example*

- Sarah Elliott's code is called MESS
<https://tcg.cse.anl.gov/papr/codes/mess.html>
- Together with TAU and working the ALCF catalyst team led by Murat Keceli and Tim Williams, Sarah reduced the total runtime of her application by a factor of **3.8x** in one day - on Oct 12, 2023.
- Sarah is a domain scientist in the Chemical Sciences and Engineering (CSE) group (not affiliated with ALCF).

Source: Sameer Shende



Sarah Elliott from Argonne – <https://tcg.cse.anl.gov/papr/codes/automech/contributor/elliott.html>.