



InPEX, Niteroi, Brazil

Navigating the Post-Exascale Computing Era: Roadmap, Energy Efficiency, Heterogeneous Computing

Jeffrey S. Vetter

Section Head, Advanced Computing Systems Research



U.S. DEPARTMENT
of **ENERGY**

ORNL IS MANAGED BY UT-BATTELLE LLC
FOR THE US DEPARTMENT OF ENERGY

Executive Summary

ORNL Roadmap

Energy Efficiency or The Gigawatt Imperative

Heterogeneous Computing Benchmarks

FY 2026 – FY 2030

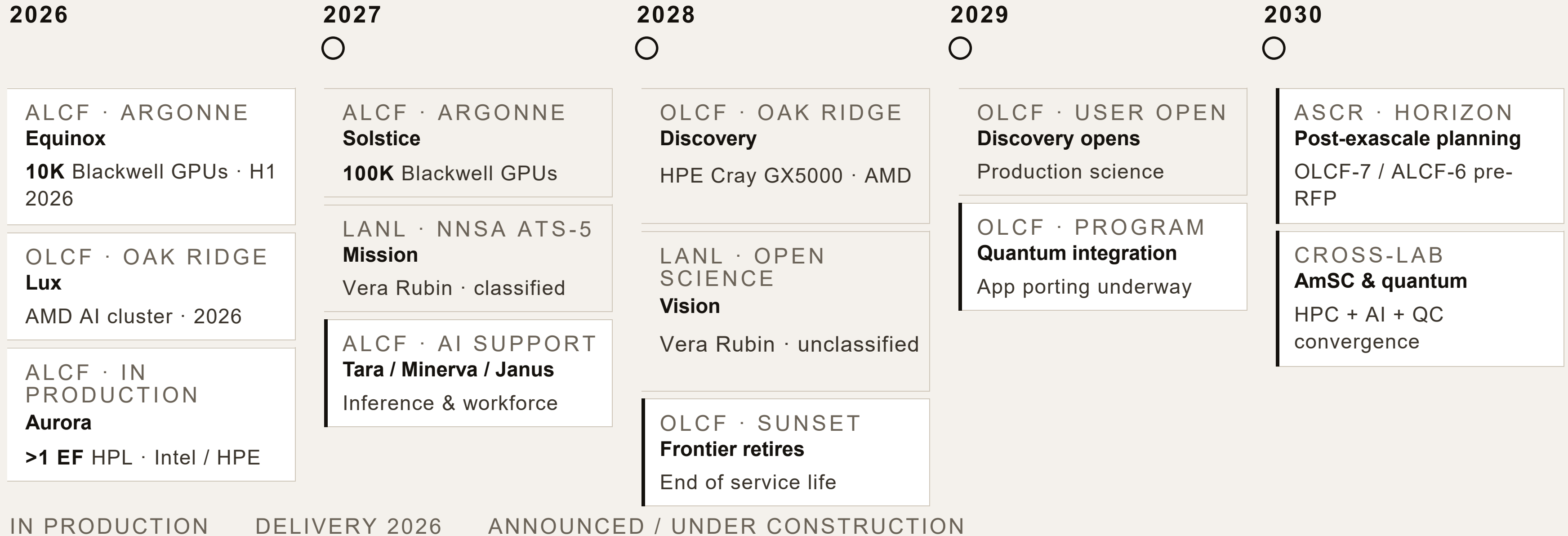
DOE HPC Roadmap, 2026–2030.

The post-exascale era is now a **procurement pipeline**. Over the next five years, the Office of Science stands up a generation of systems that fuse classical HPC with frontier-scale AI — anchored at **Argonne (ALCF)** and **Oak Ridge (OLCF)**, with NNSA classified capacity at Los Alamos.

<p>EXASCALE DEPLOYED</p> <p>3 systems</p> <p>Frontier · Aurora · El Capitan</p>	<p>AI PERF. INCOMING (ANL)</p> <p>2,200 <small>EF</small></p> <p>Solstice + Equinox, combined</p>
<p>FLAGSHIP ANL DELIVERY</p> <p>2026</p> <p>Equinox — 10,000 Blackwell GPUs</p>	<p>FLAGSHIP ORNL DELIVERY</p> <p>2028</p> <p>Discovery — 3–5× Frontier</p>

Five-year timeline.

Exascale flagships remain in production across the window. AI-optimized systems fill 2026–2027; the next leadership-class HPC system, Discovery, lands at ORNL in 2028.



Sources: nvidianews.nvidia.com/news/nvidia-oracle-us-department-of-energy-ai-supercomputer-scientific-discovery · hpcwire.com/off-the-wire/doe-seeks-proposals-for-discovery-successor-to-ornls-frontier-supercomputer · anl.gov/article/argonne-releases-aurora-exascale-supercomputer-to-researchers

The two flagship bets.

Argonne doubles down on NVIDIA-based AI at unprecedented GPU count; Oak Ridge stays on the AMD + HPE Cray lineage with a classical-HPC leadership machine sized to replace Frontier.

ANL · ALCF

ARGONNE LEADERSHIP COMPUTING FACILITY

Solstice+Equinox

DOE's largest AI supercomputer — NVIDIA + Oracle partnership.

2,200

EF AI
COMBINED

110K

BLACKWELL
GPUS

SOLSTICE	100K Blackwell GPUs — >2× El Capitan.
EQUINOX	10K Blackwell GPUs · H1 2026.
FABRIC	NVIDIA networking links both as one AI complex.
STACK	Megatron-Core training; TensorRT inference.
PARTNERS	DOE · NVIDIA · Oracle (OCI).

OAK RIDGE LEADERSHIP COMPUTING FACILITY · OLCF-6

Discovery

Successor to Frontier — HPE + AMD on Cray GX5000.

3–5×

FRONTIER
THROUGHPUT

\$500M

BUDGET

CPU	AMD EPYC " Venice " (next-gen).
GPU	AMD Instinct MI430X .
PLATFORM	HPE Cray Supercomputing GX5000 .
STORAGE	HPE Cray K3000 — factory-built DAOS.
SCHEDULE	Delivery 2028; users 2029 post-CAAR.

**ENERGY EFFICIENCY IS THE NEW
PRIORITY!**

THE SETUP

The binding constraint on AI progress has shifted from algorithms to power.

4×

per year, 2020–2025

Growth in compute for frontier AI training.

<1%

per year

Growth in U.S. electrical-grid capacity over the same period.

60–75

GW new data-center capacity, within a decade

Comparable in scale to early-20th-century national electrification.

A NEW FIGURE OF MERIT

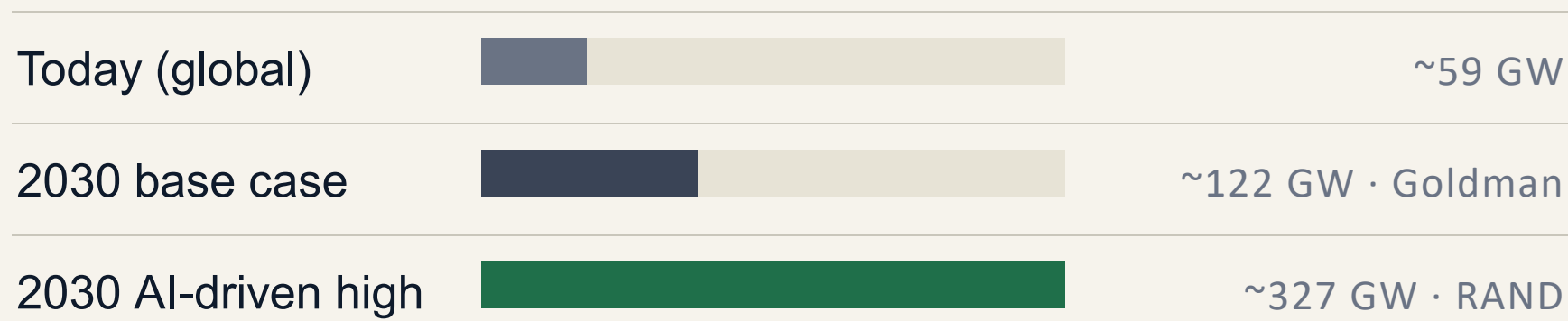
“We don't buy servers anymore. We buy power, and then we figure out what to do with it.”

INDUSTRY EXECUTIVE, OCP GLOBAL SUMMIT 2025

The appropriate figure of merit for this era is not FLOPS. It is tokens per watt per dollar — a composite spanning power, silicon, thermal, and capital efficiency.

FORECASTS REVISED UPWARD THREE YEARS RUNNING

Demand is colliding with a nearly flat grid.



Projected global data-center capacity. Range spans a factor of three.

Transformer lead times now exceed three years . Grid-interconnection queues in key U.S. markets run four to seven years.

Input cost trajectories are diverging: silicon is deflating, but transformers, tier-1 land, critical minerals, and electricity are running at 2–3× their 2023 levels.

The economic center of gravity of AI is shifting from silicon to steel, copper, and concrete.

CASE IN POINT

xAI Colossus — 1.0++ GW

Hyperscale AI campuses are already planning at the gigawatt scale — pulling nuclear-plant-sized loads out of an interconnection queue that takes longer to clear than a PhD.

ORNL's Frontier runs at roughly ~10× the performance and ~50% more power of previous ORNL flagship Summit system.



TRANSITION

DOE ASCR convened the community to define a response.

Workshop on Energy-Efficient Computing for Science — September 9–11, 2024. Alongside companion workshops on analog and neuromorphic computing.

Experts from academia, government, and industry examined the problem across eight breakouts: algorithms, hardware, data management, modeling & simulation, facility-to-edge, resource management, programming systems, and crosscuts.

Basic Research Needs in Energy-Efficient Computing for Science Published Jan 2026 · DOI 10.2172/2476961



8

breakout topic areas

5

Priority Research Directions

6

cross-cutting enablers

THE ANSWER, IN FIVE PARTS

Five priority research directions.

PRD 1 Co-design energy-efficient hardware devices and architectures.
Analog · stochastic · optical · cryogenic · neuromorphic · quantum · biological.

PRD 2 Define the algorithmic foundations of energy-efficient scientific computing.
A mathematical notation that includes the energy cost of data movement.

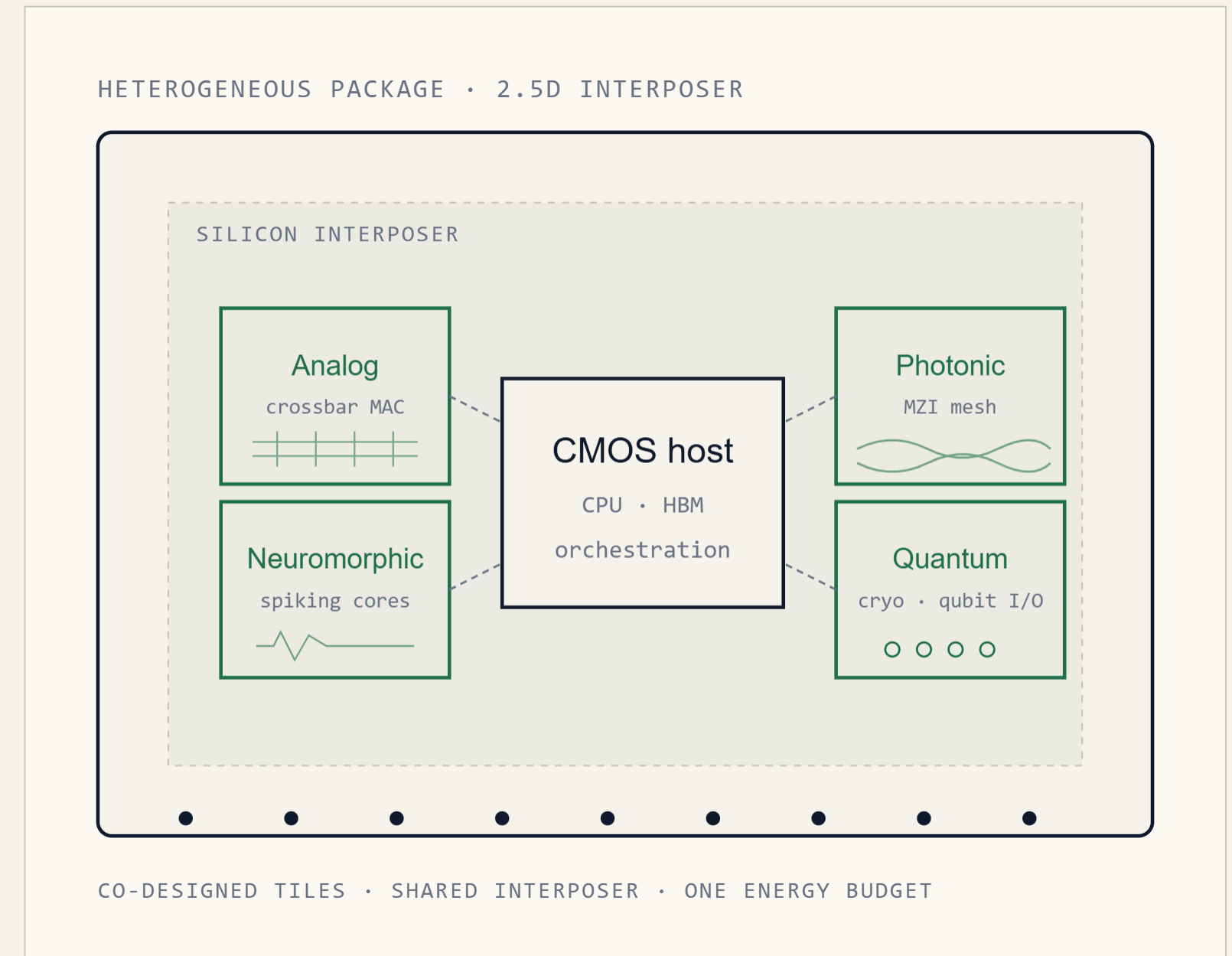
PRD 3 Reconceptualize software ecosystems for energy efficiency.
Languages, compilers, libraries, runtimes, and OS — every layer, energy-aware.

PRD 4 Enable energy-efficient data management.
Data movement is the dominant energy cost. Storage systems are still energy-agnostic.

PRD 5 Integrated, scalable energy measurement and modeling.
Device to facility. Open interfaces. Real-time, reproducible.

Co-designed devices for the workloads that matter to science.

- Take a broad view of paradigms — analog, stochastic, optical, optical, cryogenic, neuromorphic, quantum, biological.
- Evaluate end-to-end: performance, energy, manufacturing readiness, yield.
- Target the kernels that most benefit DOE workloads — then integrate specialization through chipllets and advanced packaging.
- Lower the cost of heterogeneous integration with shared chip-design software and fab access.

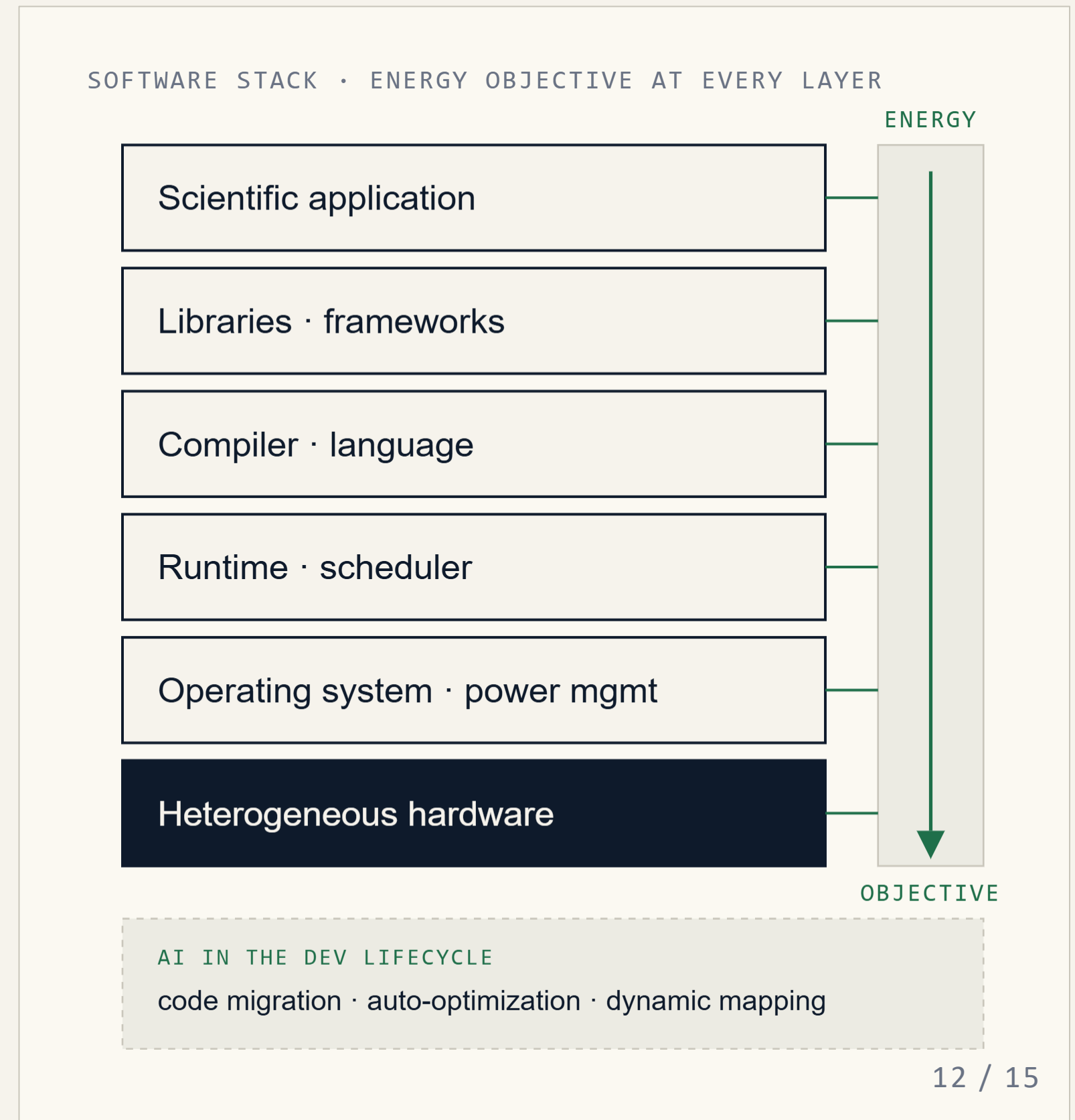


Ties to afternoon breakouts on analog, photonic, quantum, and neuromorphic systems.

A software stack where every layer is energy-aware.

- Languages and compilers that emit energy-efficient code by default.
- Libraries that hide hardware complexity without surrendering efficiency.
- Runtimes that distribute work with a power objective — not only a performance objective.
- Operating systems that arbitrate energy against performance, security, and reliability.

Embed AI into the development lifecycle — code migration, optimization, dynamic resource mapping — so every layer carries an energy objective down to silicon.



One code, many machines – the portability gap

Programming models differ in which hardware they target natively, where they run portably, and where performance must be rewritten to recover.

MODEL × TARGET – NATIVE SUPPORT & ACHIEVABLE PERFORMANCE

	CPU X86 · ARM	NVIDIA GPU CUDA	AMD GPU ROCm	Intel GPU ONEAPI	FPGA XILINX · ALTERA
Python NumPy / JAX / Numba	portable	via JAX/CuPy	emerging	emerging	none
C++ std::par / stdexec	native	via nvc++	limited	limited	none
Fortran do concurrent	native	nvfortran	partial	partial	none
CUDA	none	native	via HIPIFY	none	none
HIP	none	portable	native	none	none
OpenMP target offload	native	offload	offload	offload	none
OpenACC	fallback	native	partial	partial	none
SYCL DPC++ / AdaptiveCpp	portable	portable	portable	native	FPGA flow
OpenCL	portable	legacy	portable	portable	portable
Kokkos backend abstraction	portable	portable	portable	portable	none

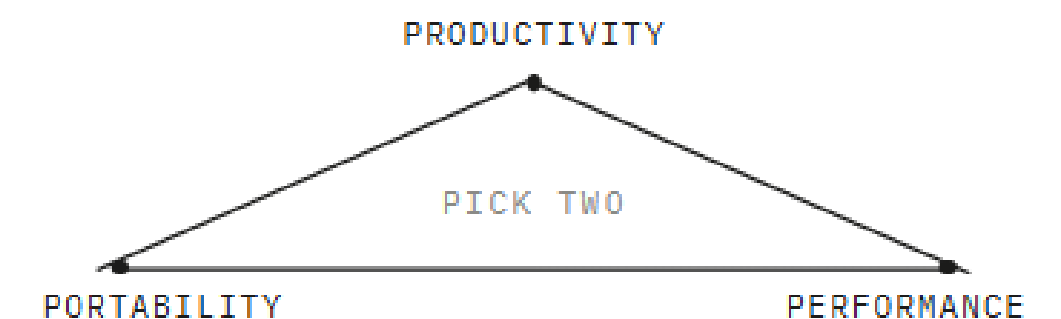
LEGEND

- **native** designed-for target, peak perf
- **portable** runs, near-native perf typical
- **partial** runs, perf gap or feature gap
- **limited** transpile or vendor-specific path
- **none** no practical path

CORE CHALLENGES

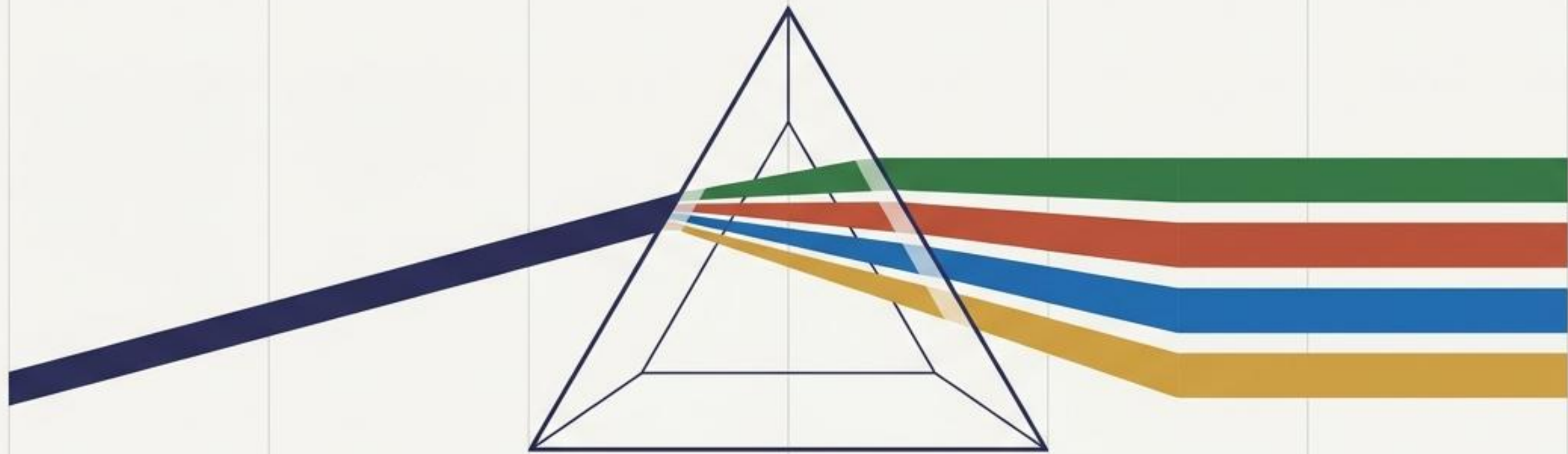
- 01 **Divergent memory models** – host/device, unified, scratchpad, HBM tiers.
- 02 **Vendor lock-in** – CUDA and ROCm siloes fragment kernel code.
- 03 **Parallelism granularity** – SIMD, SIMT, task, dataflow don't map 1:1.
- 04 **Toolchain churn** – compilers, runtimes, and ABIs drift per platform.
- 05 **Tuning cost** – portable ≠ performant; each target needs its own sweep.

THE TRADE - OFF



HeCBench: The Rosetta Stone of Heterogeneous Computing

Evaluating Performance, Portability, and Productivity Across Modern GPU Architectures



```
// A comprehensive benchmark suite written in CUDA, HIP, SYCL, and OpenMP.
```





<https://bit.ly/hecbench>

>5,400 Commits

300+ Benchmark Variations

19+ Scientific Domains

The HeCBench Portability Matrix

	CUDA	Native execution mapped strictly to NVIDIA architectures.
	HIP	AMD's native C++ runtime; heavily evaluated for NVIDIA cross-compatibility.
	SYCL / DPC++	Intel's native model; evaluated for CPU/GPU/FPGA portability via LLVM.
	OpenMP 4.5+	Evaluated for pragma-based target offloading capabilities.

Data Version Control (DVC)

Seamless provisioning of massive external benchmark datasets.

Unified Build Scaling

Central CMake architecture orchestrating 200+ distinct implementations.

Automated Verification

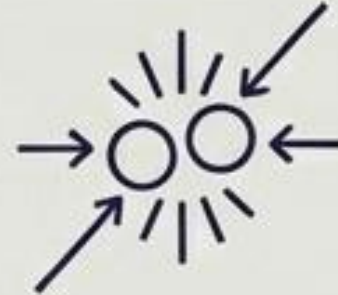
Built-in evaluation comparing host and device result accuracy.

Translating Heterogeneous Models into Scientific Reality



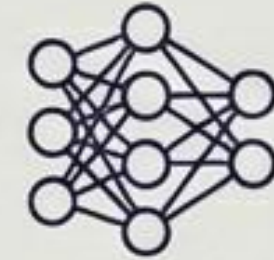
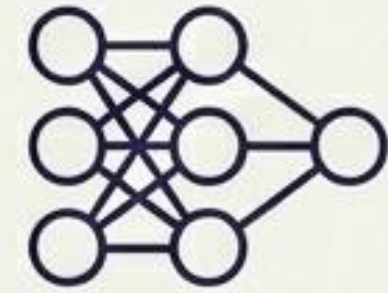
Bioinformatics

- **bsw**: GPU-accelerated Smith-Waterman algorithm for large-scale batch alignments.
- **minimap2**: Hardware-accelerated long-read pairwise overlapping, critical for modern genome sequencing.



Physics & Simulation

- **lulesh**: Livermore unstructured Lagrangian explicit shock hydrodynamics (evaluating structural responses).
- **bh**: Barnes-Hut n-body algorithm simulating complex gravitational forces within star clusters.



Machine Learning & AI

- **attention** & **moe**: Paged attention mechanisms and Mixture-and Mixture-of-Experts routing mapping modern LLM topologies to hardware.
- **adamw**: Hardware mapping of Adaptive Moment Estimation with weight decay.

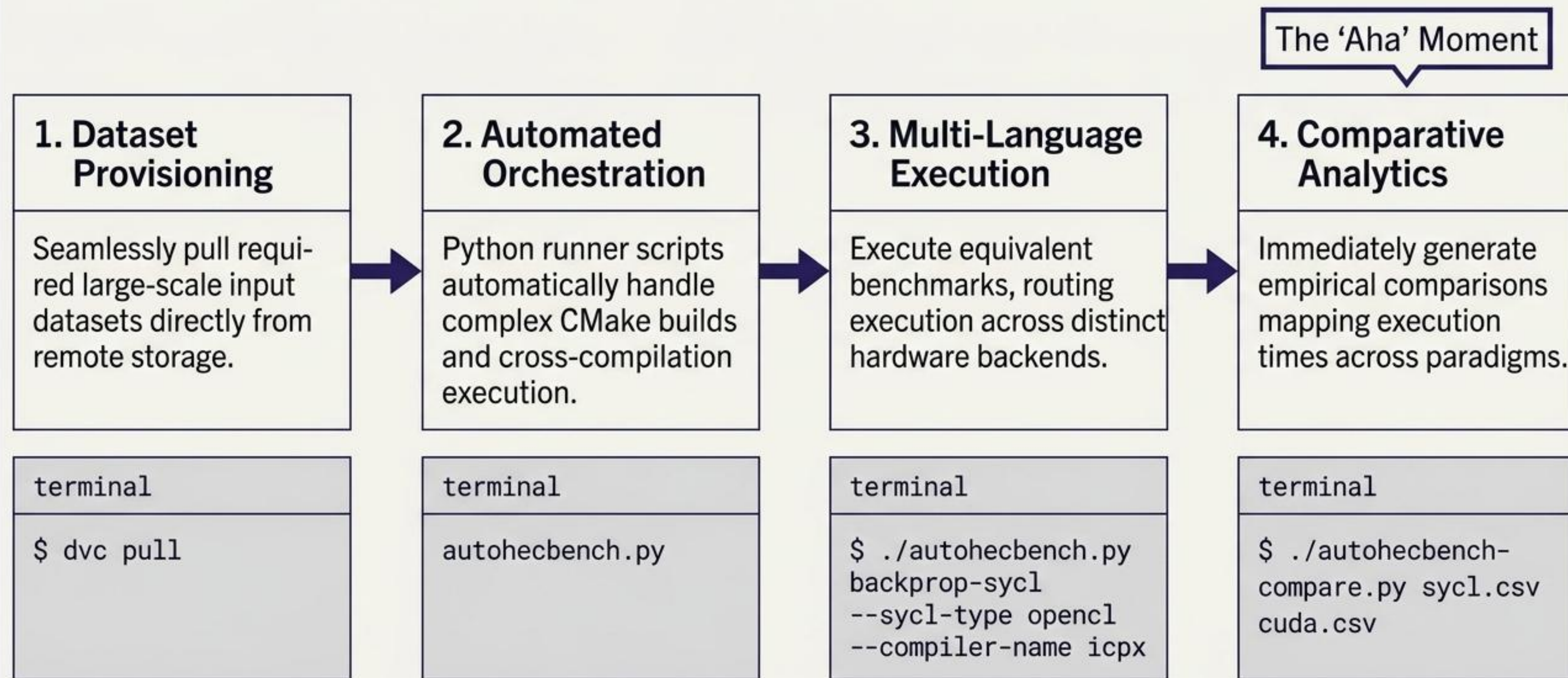
Breaking news... our RIKEN CCS colleagues are using CoPilot to port these benchmarks to Kokkos (with great success)

Porting HeCBench Benchmarks to Kokkos

Performance Comparison: Kokkos (OpenMP) vs OMP Target on ARM aarch64 CPU

System: ARM aarch64 · 20 cores @ 2808 MHz · Ubuntu 22.04 · Kokkos 3.7.01

The Practitioner's Workflow: Code to Insight



Summary

ORNL Roadmap

Energy Efficiency or The Gigawatt Imperative

Heterogeneous Computing Benchmarks

BONUS