STICM – Smart Tools and Interfaces for Complex Memories

Scalable Tools Workshop 2026, July 27 2026

Terry Jones, ORNL
Michael R. Jantz, UTK
J. Zach McMichael, UTK

U.S. DEPARTMENT OF ENERGY

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

Introducing STICM

- STICM: Smart Tools and Interfaces for Complex Memory
  - Joining the STEP portfolio in FY 2026-2027
  - Will steward and extend the high-level interface of the DOE ECP project SICM
- Today's talk
  - Goals and design of SICM
  - Major contributions of SICM
  - Ongoing and planned work for STICM

STICM
SMART TOOLS AND INTERFACES FOR COMPLEX MEMORIES

Terry Jones, ORNL
PI

Mike Jantz, UTK
Co-PI

Zach McMichael, UTK
PhD Candidate

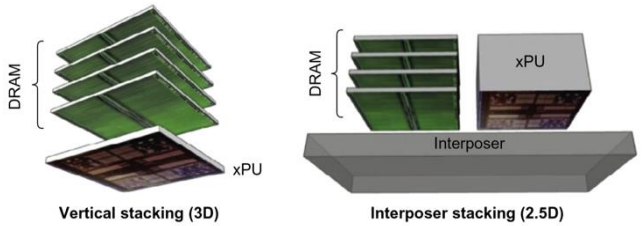
David Rabson, DOE
Program Manager

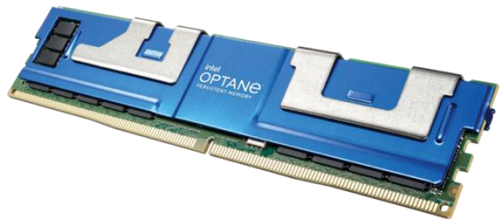
STICM is supported by these institutions

Rise of Complex Memories

- **Trends placing new strains on memory systems**
  - Success and proliferation of AI, data analytics
  - “Scaling up” memory has become more difficult
  - Increasing demands for high density sharing in HPC

- **New media technologies**



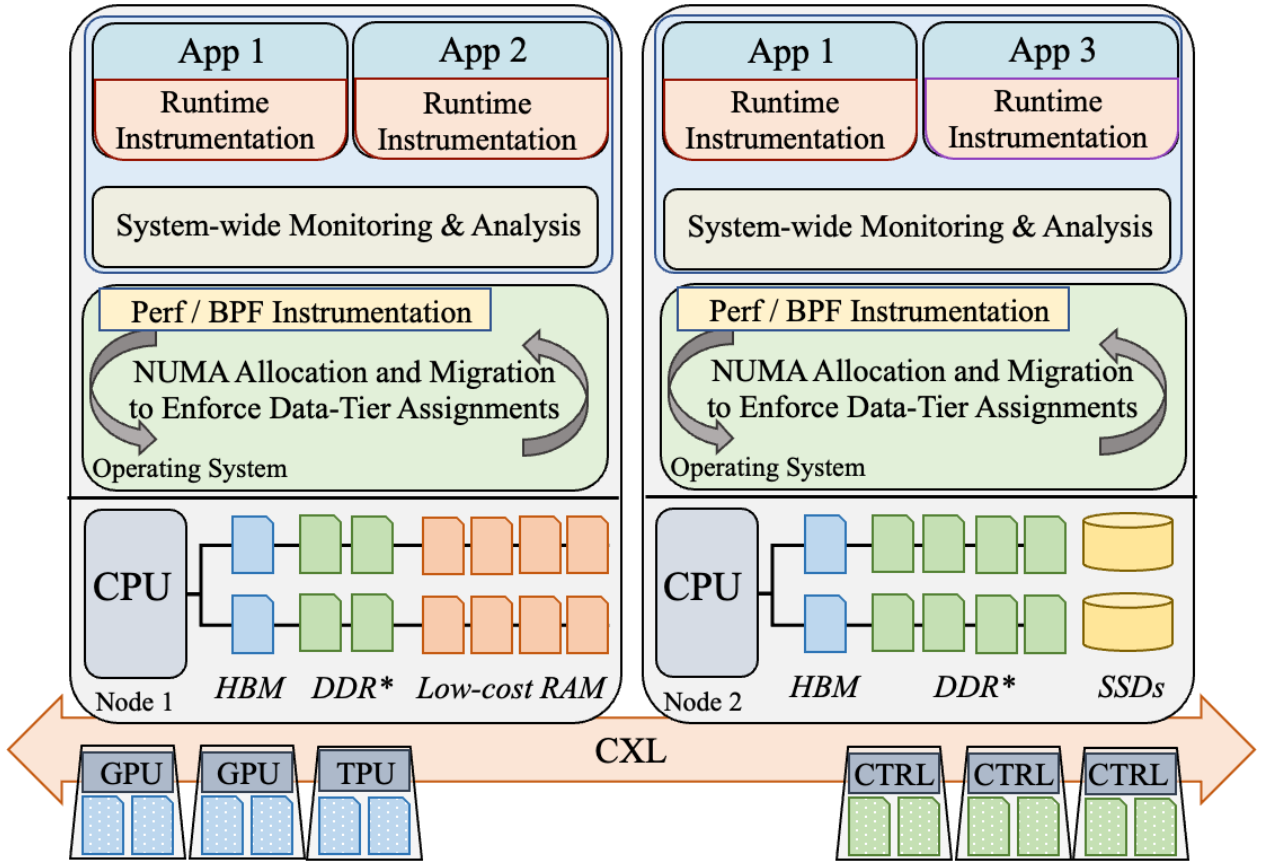


CXL Compute Express Link

How to Manage Complex Memories

- **Application behaviors are very diverse**
  - Difficult to create a one-size-fits-all solution
  - Solutions should be tailored to individual applications
- **Applications are not used to managing memory HW**
  - Memory intensive applications are extremely complex
  - Solutions need to be simple, transparent, portable

SICM Overview



- High-level interface
  - Tools / strategies to collect and use mem mgmt guidance from running apps

Terry Jones, ORNL
Mike Jantz, UTK
Ben Olson, UTK
Brandon Kammerdiener, UTK
Zach McMichael, UTK

- Low-level interface
  - Portable API based on existing Linux infrastructure

Michael Lang, LANL
Latchesar Ionkov, LANL
Scott Pakin, LANL
Jason Lee, LANL
Terry Turton, LANL
Roxana Bujack, LANL
Li Tang, LANL

- Specialized Graph techniques
  - Includes Metall API to allocate custom C++ data structures

Maya Gokhale, LLNL
Keita Iwabuchi, LLNL

Bujack, R., Gokhale, M., Ionkov, L., Iwabuchi, K., Jantz, M., Jones, T., Lakshmiranganatha, S., Lang, M.K., Lee, J., Olson, M.B. and Pakin, S., 2025. The ECP SICM project: Managing complex memory hierarchies for exascale applications. *The International Journal of High Performance Computing Applications*, *39*(1), pp.193-207.

SICM High-Level Contributions

- **Guided Memory Management with SICM**
  - Evaluating the Effectiveness of Program Data Features for Guiding Memory Management [1, 2]
  - MemBrain: Automated Application Guidance for Complex Memory Systems [3, 4, 5, 6]
  - Flexible and Effective Object Tiering with MAT Daemon [7, 8]

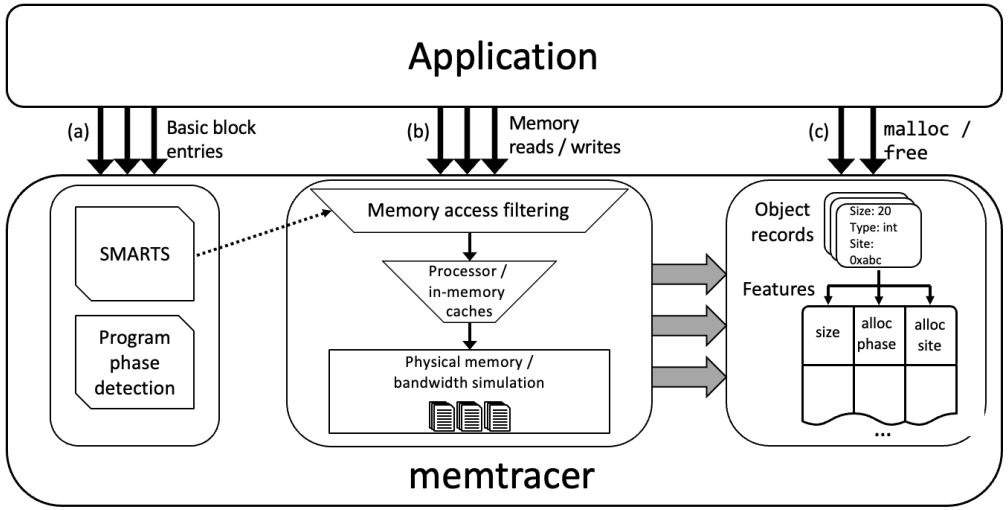
Guiding Memory Management with Data Features

- Understanding and predicting memory usage is hard
- Common strategy:
  - Combine and associate memory usage profiles with some *program data features*
  - Examples: object size, type, or allocation context
  - If features accurately distinguish data with similar usage, profiling can be more efficient
- Problem: which program features are most effective for guiding memory management?

Effler, et al. Evaluating the Effectiveness of Program Data Features for Guiding Memory Management. MemSys 2019.

Methodology

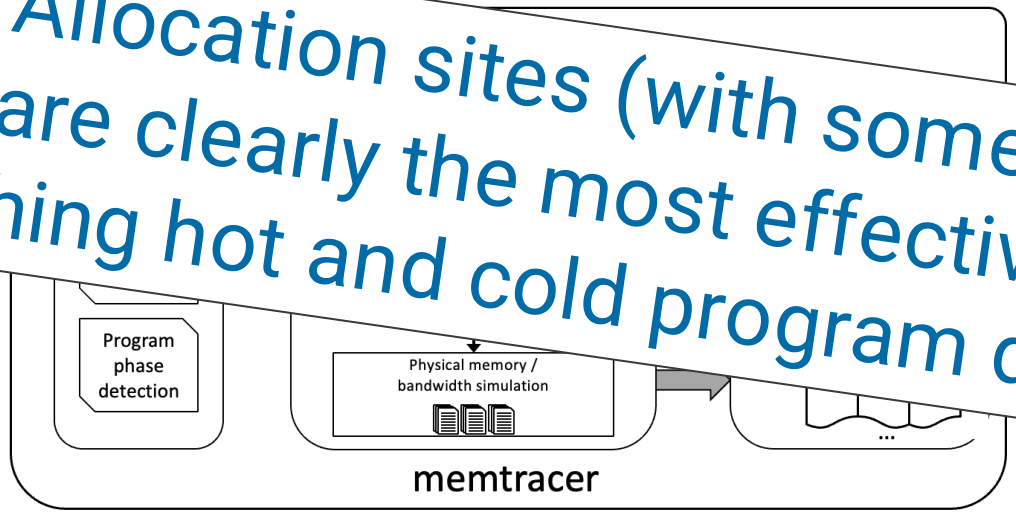
- Designed a simulation framework to:
  - Profile the usage of individual data objects
  - Associate and group object profiles with common data features
    - Application
    - Size
    - Size bucket
    - Type
    - Allocation phase
    - Alive phases
    - Access signature
    - Allocation site
    - Allocation context



Methodology

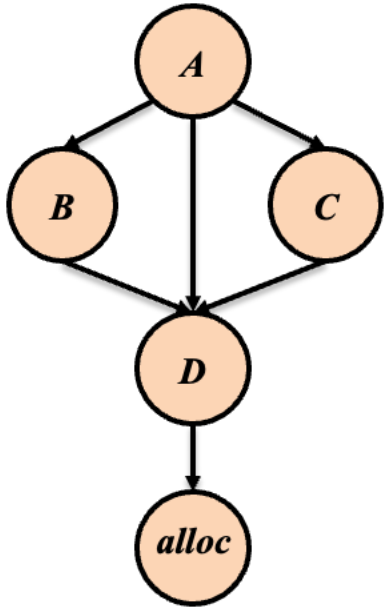
- Designed a simulation framework to:
  - Profile the usage of individual data objects
  - Associate and group object profiles with common data features
    - Allocation phase
    - Alive phases
    - Access signature
    - Allocation site
    - Allocation context

Primary Finding: Allocation sites (with some call stack context) are clearly the most effective feature for distinguishing hot and cold program data



Automated Application Guidance for Complex Memories

- Track usage for *program allocation sites*
  - Site = allocation instruction + call path
  - Intuition 1: origin can help predict usage
  - Intuition 2: profile transferability



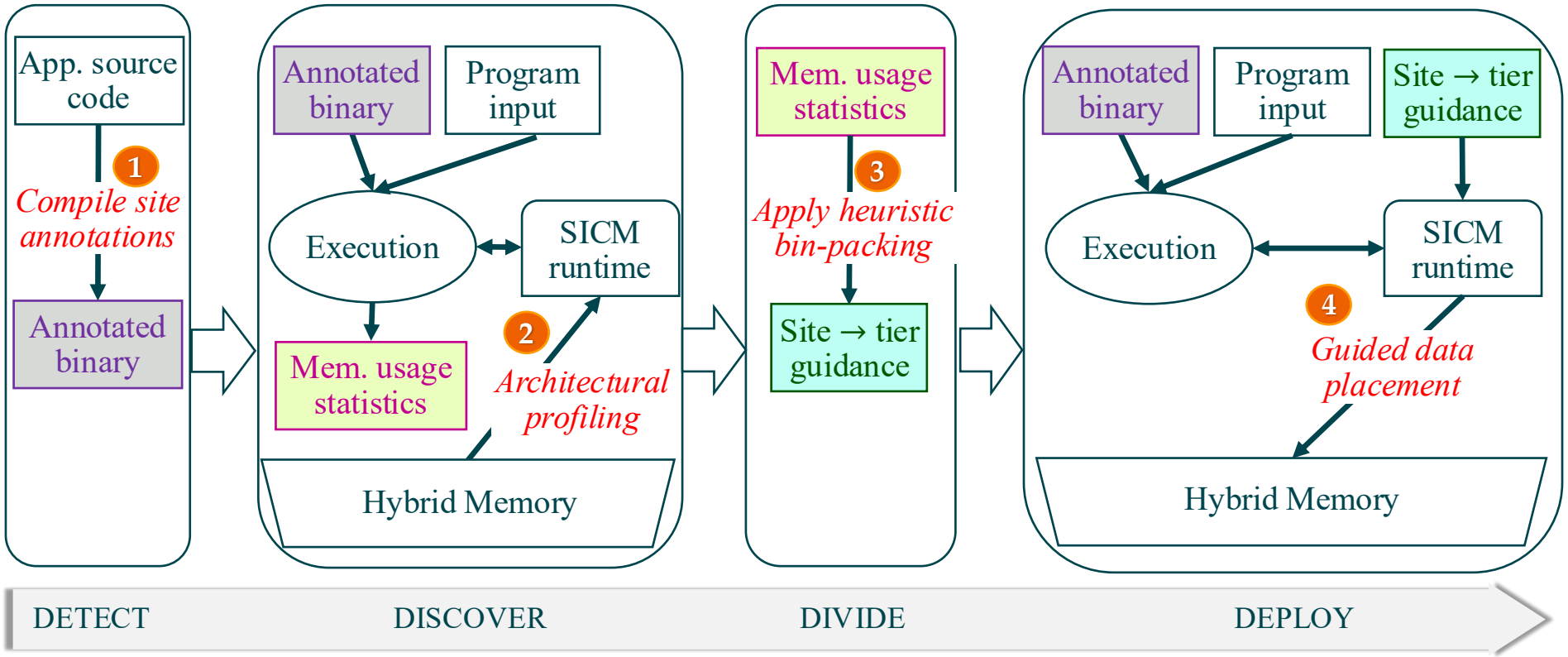
Potential Sites

A B D

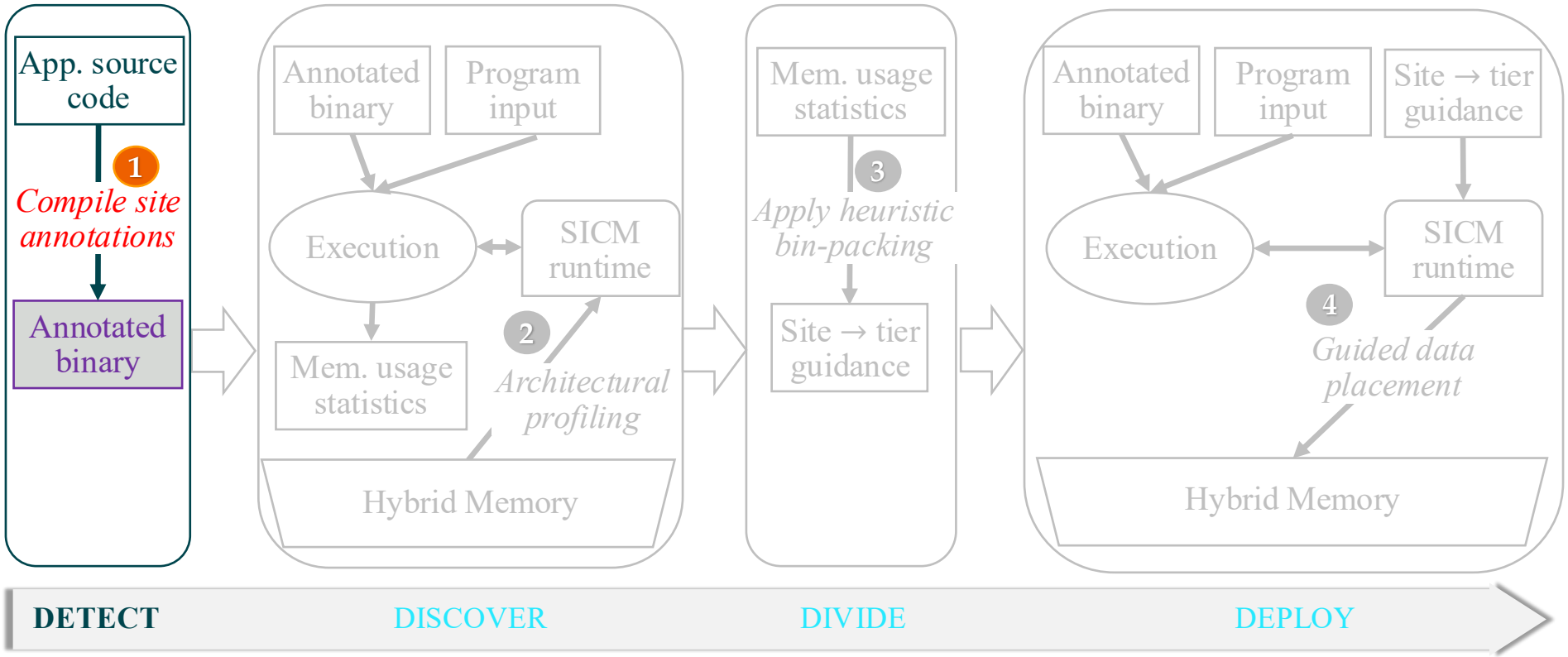
A D

A C D

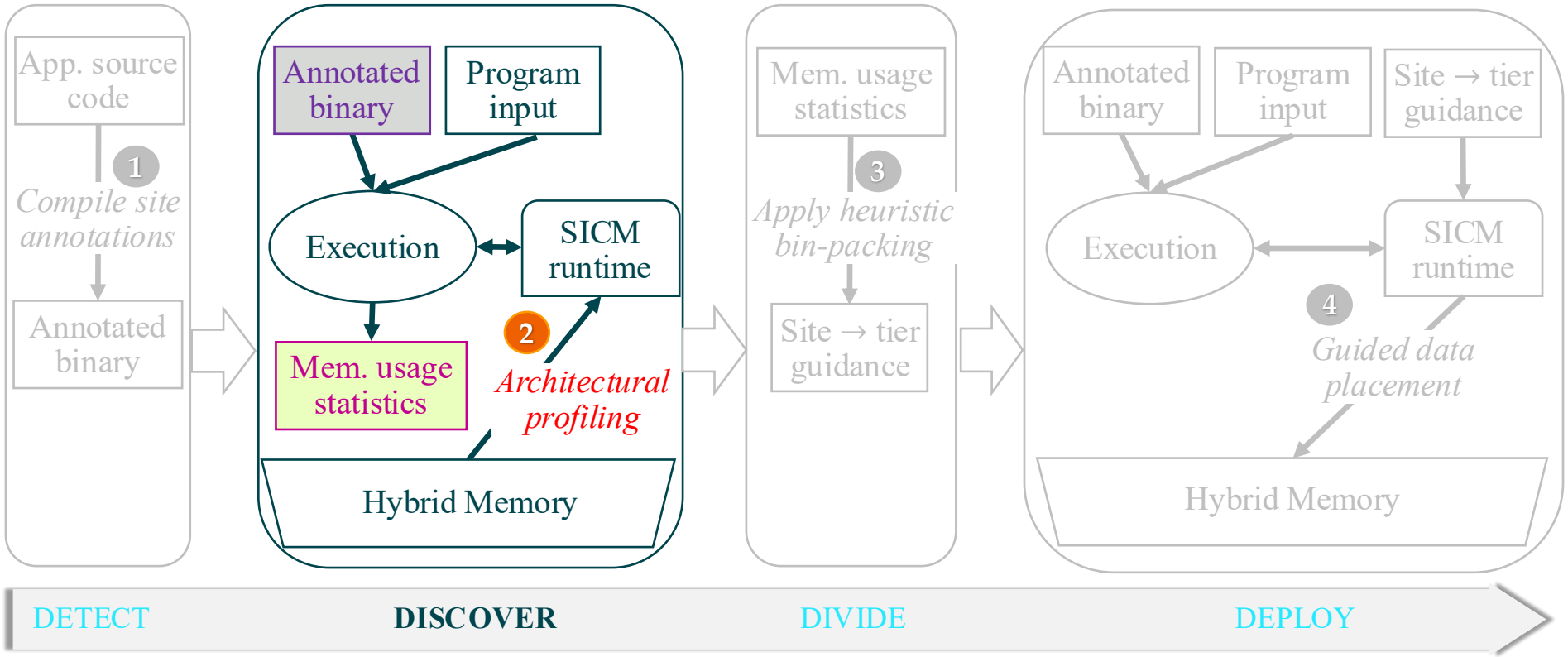
SICM High-Level Overview



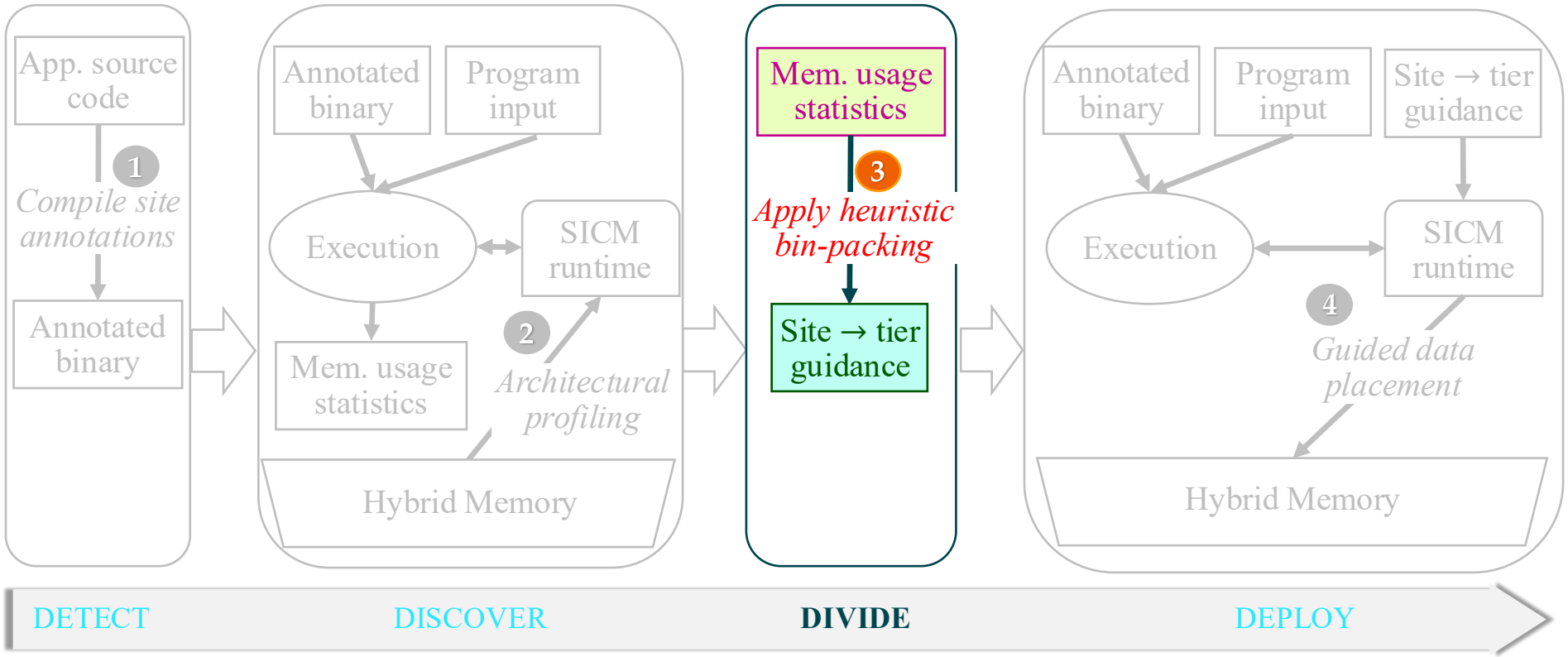
SICM High-Level Overview



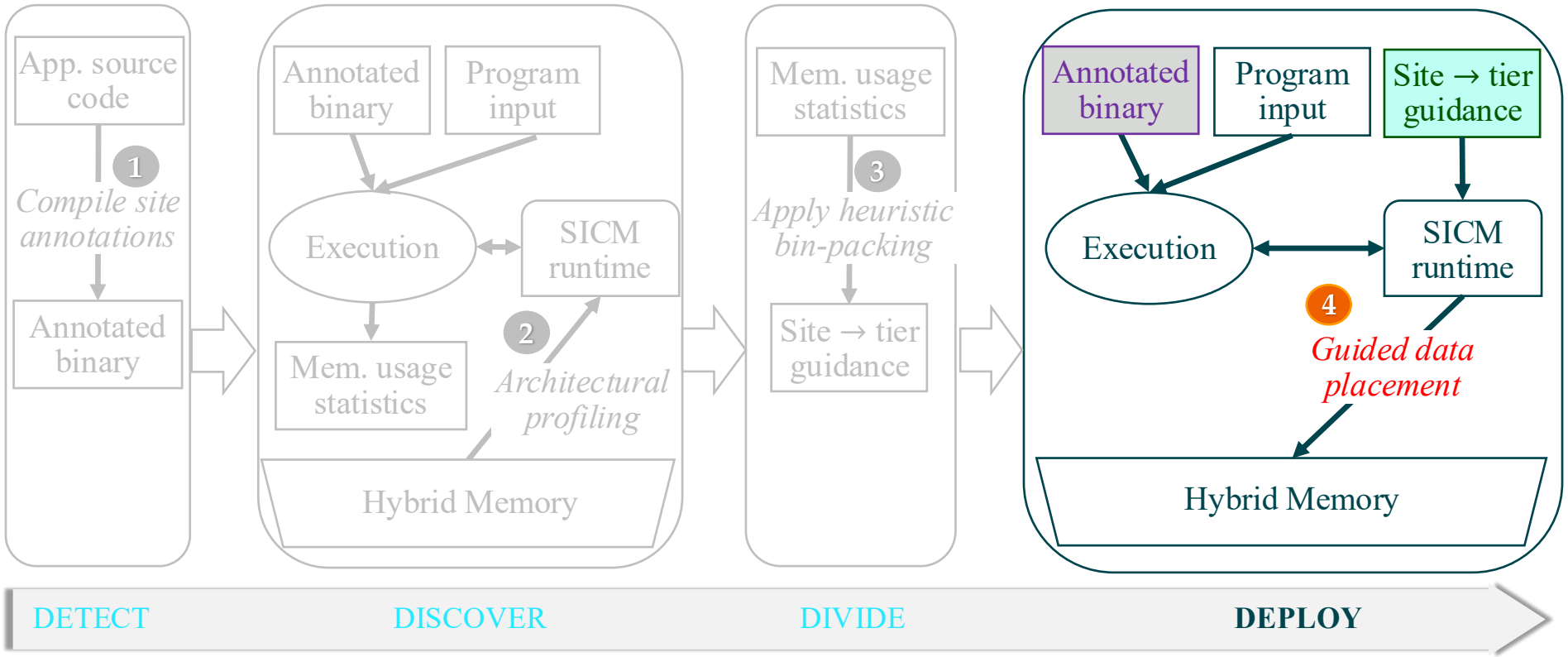
SICM High-Level Overview



SICM High-Level Overview

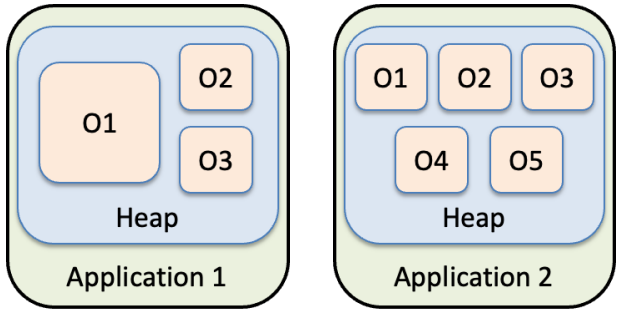


SICM High-Level Overview



Object Tiering for Complex Memories

- Drawbacks to MemBrain
  - Requires recompilation, offline profiling
  - Only supports single-process workloads
- New approach: track usage for *individual data objects*
- Two major design components
  - Custom allocator to track object creation / removal
  - Process (MAT Daemon) to aggregate, analyze object profiles, calculate and enforce new data tiering assignments



Going Forward with *STICM*

- Adapting Object Tiering for New Architectures
  - High-bandwidth memories (SPR)
  - More diverse architectures connected via CXL

- Exploit deeper integration with OS / hardware
  - Coordinate physical mem mgmt w/ object tiering
  - Hardware-accelerated data migration

- Opportunities in high-level language applications
  - Leverage HLL-VM features to reduce fast mem capacity

References

1. T. Chad Effler, Brandon Kammerdiener, Michael R. Jantz, Saikat Sengupta, Prasad A. Kulkarni, Kshitij A. Doshi, and Terry Jones. Evaluating the Effectiveness of Program Data Features for Guiding Memory Management. In the *ACM/IEEE International Symposium on Memory Systems (MemSys '19)*, October 2019. [pdf]
2. T. Chad Effler, Adam P. Howard, Tong Zhou, Michael R. Jantz, Kshitij A. Doshi, and Prasad A. Kulkarni. On Automated Feedback-Driven Data Placement in Hybrid Memories. In the *International Conference on Architecture of Computing Systems (ARCS '18)*, ser. Lecture Notes in Computer Science. April 2018. [pdf, slides]
3. M. Ben Olson, Tong Zhou, Michael R. Jantz, Kshitij A. Doshi, M. Graham Lopez, and Oscar Hernandez. MemBrain: Automated Application Guidance for Hybrid Memory Systems. In the *13th IEEE International Conference on Networking, Architecture, and Storage (NAS '18)*. **Awarded Best Paper.** October 2018. [pdf]
4. M. Ben Olson, Brandon Kammerdiener, Michael R. Jantz, Kshitij A. Doshi, and Terry Jones. Portable Application Guidance for Complex Memory Systems. In the *ACM/IEEE International Symposium on Memory Systems (MemSys '19)*, October 2019. [pdf]
5. M. Ben Olson, Brandon Kammerdiener, Michael R. Jantz, Kshitij A. Doshi, and Terry Jones. Online Application Guidance for Heterogeneous Memory Systems. In the *ACM Transactions on Architecture and Code Optimization (TACO)*, April 2022. [pdf]
6. T. Chad Effler, Michael R. Jantz, and Terry Jones. Performance Potential of Mixed Data Management Modes for Heterogeneous Memory Systems. In the *IEEE Workshop on Memory Centric High Performance Computing (MCHPC)*, November 2020. [pdf]
7. Brandon Kammerdiener, J. Zach McMichael, Michael R. Jantz, Kshitij A. Doshi, and Terry Jones. Flexible and Effective Object Tiering for Heterogeneous Memory Systems. In the *ACM SIGPLAN International Symposium on Memory Management (ISMM '23)*, June 2023. [pdf]
8. Brandon Kammerdiener, J. Zach McMichael, Michael R. Jantz, Kshitij A. Doshi, and Terry Jones. Flexible and Effective Object Tiering for Heterogeneous Memory Systems. In review at ACM Transactions on Architecture and Code Optimization *(ACM TACO)*, May 2024.

I’m happy to respond to any QUESTIONS
My email address is **trj@ornl.gov**

this presentation – Scan Me

STICM

SMART TOOLS AND INTERFACES
FOR COMPLEX MEMORIES

Notice: This research was supported by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration. The presentation has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a non-exclusive, paid up, irrevocable, world-wide license to publish or reproduce the published form of the manuscript, or allow others to do so, for U.S. Government purposes. The DOE will provide public access to these results in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

U.S. DEPARTMENT OF ENERGY

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

OAK RIDGE National Laboratory