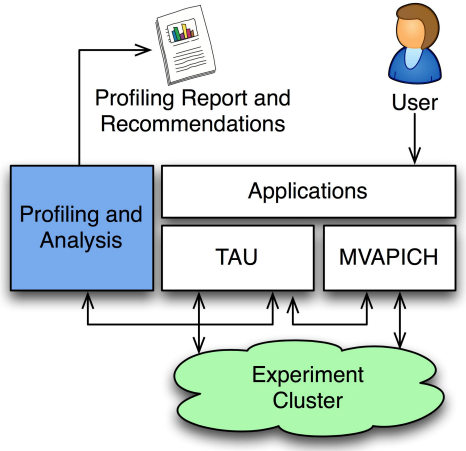
SI2-SSI:Collaborative Research: A Software Infrastructure for MPI Performance Engineering: Integrating MVAPICH and TAU via the MPI Tools Interface

Co-PIs: D. K. Panda, Sameer Shende
Institutions:Ohio State University, University of Oregon

Award #: ACI-1450440
& ACI-1450471

Research Challenges

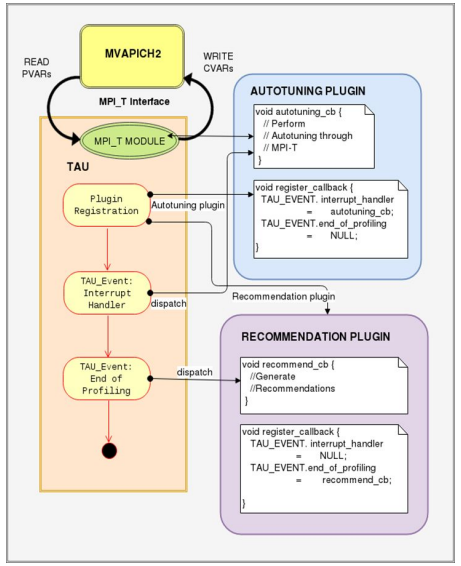
Creating an MPI programming infrastructure that can integrate performance analysis capabilities more directly, through the MPI Tools Information Interface, monitor Performance metrics during run time, and deliver greater optimization opportunities for scientific applications.



Enabling Runtime Control

- TAU defines a plugin API to deliver access control to the internal plugin map
- User can specify a regular expression to control plugins executed for a class of named states at runtime

Plugin Infrastructure



- *Fully-customizable* plugin infrastructure based on callback event handler registration for salient states inside TAU:
  - Function Registration / Entry / Exit
  - Phase Entry / Exit
  - Atomic Event Registration / Trigger
  - Init / Finalize Profiling
  - Interrupt Handler
  - *MPI_T*
- Application can define its own "trigger" states and associated plugins

Enhanced MPI_T Support in MVAPICH2

- Added MPI_T PVARs for various MPI collectives (Bcast, Reduce, Allreduce etc) to measure
- Added PVARs and CVARs for host and device based MPI operations
- Added MPI_T PVAR timers to measure the time taken and the number of calls pertaining to various collective algorithms (allreduce, barrier, reduce etc)
- Added support for dynamic MPI_T PVAR counter arrays where each index in the array represents a counter for a "bucket" or a user specified message range
- Added new CVARs that can be tuned at **run-time**

Phase-based Recommendation

- MiniAMR: Benefits from hardware offloading using SHArP hardware offload protocol supported by MVAPICH2 for MPI_Allreduce operation
- Recommendation Plugin:
  - Registers callback for *"Phase Exit"* event
  - Monitors message size through PMPI interface
  - If message size is low and execution time inside MPI_Allreduce is significant, a recommendation is generated on ParaProf (TAU's GUI) for the user to set the CVAR enabling SHArP

Per-thread, Per-phase Recommendation Generated as Metadata on ParaProf