Diamond Agent: Federated HPC Resources as a Service for Agents

Haotian Xie[*], *Junlin Chen*[*], *Mingkai Zheng*[*], *Zhao Zhang*[*]
[*]*Rutgers University*

Motivation

Agents have become capable, but their execution environments remain fragmented.

- Modern agents can independently complete complex HPC workflows, yet they are typically confined to a single supercomputer. Diamond Agent provides a persistent workspace above individual systems, allowing one agent to access multiple supercomputers while maintaining shared context.

Agents make HPC workloads portable.

- Because agents can rapidly rebuild environments and adapt applications to new systems, workloads no longer need to remain where they were initially launched. Diamond Agent exploits this mobility to migrate or resize jobs across supercomputers, avoiding congested queues and improving time-to-solution.

Diamond Agent Approach

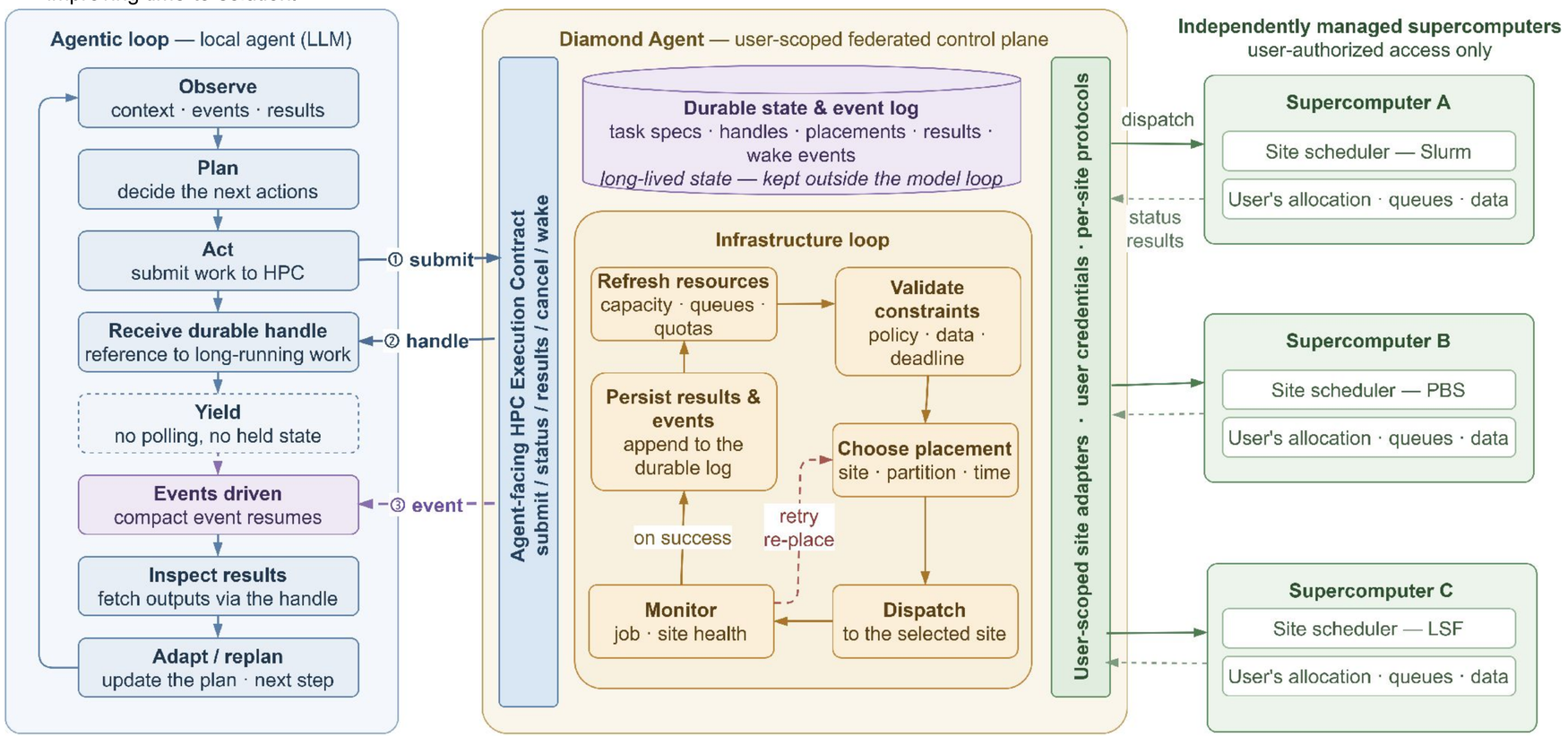
Diamond Agent provides a persistent control plane between one AI agent and multiple independently managed supercomputers.

- **One agent, one interface**. The agent runs in a local workspace and interacts with all authorized HPC systems through the same set of typed operations, instead of running a separate agent on each login node.
- **Observe and choose resources**. Diamond Agent continuously maintains a Global Resource View of available systems, accelerators, queues, and endpoint health. A **ResourceEnvelope** describes the workload's minimum and maximum useful resources and compatibility requirements, allowing Diamond Agent to select a feasible execution site from live system state.



- **Execute across supercomputers**. The persistent Task Orchestrator translates the selected action into site-specific jobs and delegates execution through Globus Compute, while Globus Transfer handles large cross-site data movement. Tasks can be retried, resized, or re-placed when conditions change.

The agent decides what to do; Diamond Agent decides where and how to execute it across the federation.



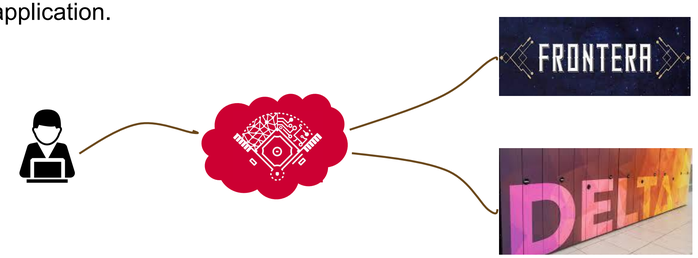
Applications

Autonomous heterogeneous execution for Terminal-Bench

Diamond Agent enables a single agent to autonomously coordinate one application across multiple HPC systems without human intervention.

- DeltaAI GPUs: deploy and serve the LLM.
- Frontera CPUs: deploy parallel sandboxes and execute Terminal-Bench tasks.
- Cross-system coordination: the agent uses Frontera SSH tunneling to connect the CPU-side sandboxes with the LLM service running on DeltaAI.

Starting from a local workspace, one agent autonomously orchestrates GPU and CPU resources across two supercomputers as a single application.



Adaptive multi-supercomputer execution

Diamond Agent enables a single agent to autonomously distribute and adapt parallel workloads across multiple supercomputers without human intervention.

- Federated execution: four GPU workers were initially distributed across Vista, Delta, DeltaAI, and Anvil.
- Queue-aware adaptation: when the Anvil task remained queued, Diamond Agent automatically re-placed it on Vista after the other workers completed.
- Failure recovery: when an execution endpoint was deliberately made unavailable, Diamond Agent detected the failure and submitted a replacement within 60 seconds.

https://github.com/Diamond-Proj/diamond-agent