IBM – Illinois Discovery Accelerator Institute
Focus Areas for Fall 2024 RFP

Table of Contents

1 CONTEXT 2

2 INTRODUCTION 2

3 GENERAL OBJECTIVES 3

4 HYBRID CLOUD & AI THRUST 3

4.1 AGENTIC SYSTEMS 3
4.1.1 LLM as an Abstraction (LLMaaA) 4
4.2 AI MODEL OPTIMIZATION AND RUNTIMES 5
4.2.1 Enabling the Edge Transformation: End-to-end Edge-Cloud Deployment 6
4.3 HYBRID CLOUD PLATFORM AND MIDDLEWARE 7
4.3.1 Robustness, Dependability, and Scalability 9
4.3.2 Energy Optimization and Sustainability 9
4.4 INFRASTRUCTURE AND SECURITY 10
4.4.1 Application-Adaptive Cloud System for Dynamic AI Workloads 12

5 QUANTUM COMPUTING THRUST 12

5.1 THE PATH TOWARDS QUANTUM-CENTRIC SUPERCOMPUTING 13
5.2 QUANTUM-UTILITY ERA COLLABORATIONS 13

6 MATERIALS DISCOVERY THRUST 13

7 CLIMATE & SUSTAINABILITY THRUST 15

8 SUMMARY 16

1 Context

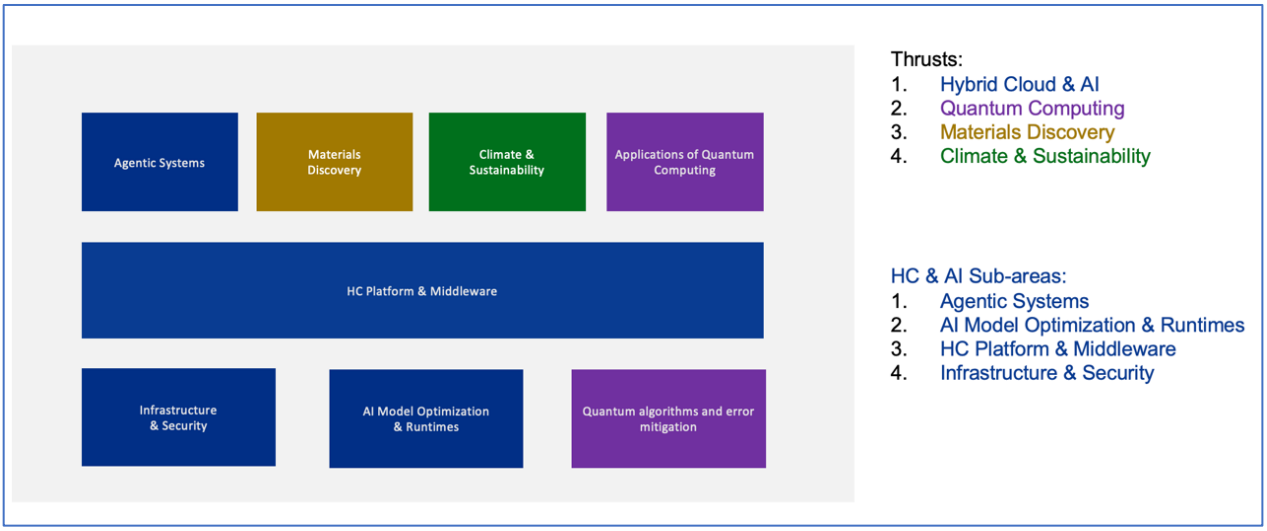
The IBM-Illinois Discovery Accelerator Institute (IIDAI) was started in 2021. This document contains a new Request for Proposals (RFP) released on November 11, 2024. This RFP is prepared by the technical leadership of the Institute from IBM Research and the University of Illinois Urbana-Champaign to identify the priority focus areas in which project proposals are being solicited for funding. The timeline, review logistics, project duration, general info session, info for thrust-level consulting office hours, and other relevant info related to the RFP are included in the announcement email of this RFP.

The Institute technical portfolio is organized into four thrusts: Hybrid Cloud & AI, Quantum Computing, Materials Discovery, and Climate & Sustainability. In addition, there are two cross-cut initiatives, one on Diversity, Equity & Inclusion, and one on Entrepreneurship. The focus areas for the crosscuts are not described in this document.

2 Introduction

The following diagram illustrates the four main thrusts mentioned above, which are covered by this RFP. In addition, the Hybrid Cloud & AI thrust is further divided into four focus sub-areas – Agentic Systems, AI Model Optimization & Runtimes, HC (Hybrid Cloud) Platform & Middleware, and Infrastructure & Security. The goals for the joint research work in each thrust and sub-area have been discussed thoroughly at the annual meeting in October 2024, at IBM T.J. Watson Research Center, Yorktown Heights, NY, as well as covered by a vision white paper, authored by both IBM Research and the University of Illinois Urbana-Champaign. This has yielded the agreed-upon high-level research topics outlined under each section below. The vision white paper can be accessed here:

2024 IIDAI Vision White Paper



3 General Objectives

The Institute encourages proposal submissions that have the following characteristics.

- Exhibit ambitious goals with well-integrated group/multi-faculty submissions in certain focus areas that may have a much larger impact than smaller projects.
- Contribute to the open source and IBM projects mentioned in the RFP.
- Span more than one thrust (e.g., Hybrid Cloud and Quantum), using applications that benefit from cross-thrust collaboration as drivers to advance innovation across thrusts, thereby highlighting the Institute's broader impact on science and society.
- Have entrepreneurship potential, especially aiming to produce successful open-source projects and/or build an open community that will advance the state-of-the-art, with concrete plans on how to achieve those goals. Entrepreneurship in this context isn't about starting a new business or creating a startup. Instead, it emphasizes fostering innovation and delivering real-world impact within industrial settings and communities.

4 Hybrid Cloud & AI Thrust

Thrust Leads: Alaa Youssef (IBM), Lav Varshney (UIUC)

Hybrid cloud refers to a computing environment where applications are run using a combination of computing, storage, and services in different environments—public clouds and private clouds, including on-premises data centers, or edge locations. Hybrid cloud allows the continued use of on-premise infrastructure while taking advantage of the options offered by public cloud providers. In this RFP, we would like to explore the future transformation of Hybrid Cloud to support new types of workloads, especially emerging AI workloads, including LLMs, agentic workflows, as well as simulation and scientific applications that are increasingly becoming AI-centric. Our focus is on ease of use, affordability, adaptability, and ubiquity at very large scales. Please refer to the vision white paper mentioned in Section 2 for more details. There are four focused sub-areas in this Thrust.

4.1 Agentic Systems

Sub-Area Leads: Avi Sil (IBM), Reyhaneh Jabbarvand (Illinois), Martin Hirzel (IBM Advisor)

Broadly, agentic workflows empower generative AI models to tackle more complex, real-world problems by providing them with a structured approach, increased autonomy, and the ability to learn and adapt.

We seek research proposals that will explore all aspects of this emerging area. Suggested high level problem domains include:

Advancing Collaboration and Intelligence in Agentic AI Systems

- Improve how AI agents collaborate, as in human collective intelligence.
- Achieve better task decomposition, chaining, and self-improvement, offering new research pathways in AI systems design.
- Overcome limitations of performing SWE and SRE tasks.

LLM-as-an-Abstraction (LLMaaA) (more details in Section 4.1.1)

- Novel paradigm for engaging with cloud computing and services, with a natural language interface.

- Interface for building, deploying, and managing complex applications.
- Driver for improving the adaptive programming model, and advancing secure, scalable cloud systems that flexibly integrate LLM and specialized agents for evolving real-world applications.

Challenges

- Significant challenges in automation, flexibility, reliability, reconfigurability, and evolvability.
- Agentic AI systems must be highly reliable to successfully achieve complex goals and take actions in complex environments.
- Need to perform better planning and reasoning.
- Ways to develop agentic FMs (Foundation Models) capable of making decisions and interacting with the world to solve long-horizon tasks through extensive interactions.
- Overcome limitations of performing SWE tasks. Enable LLMs to better understand and reason about code execution, even simulate code execution and use outcome in solving programming problems.
- Agentic systems for AI compliance and policy generation including policy as code.
- New benchmark datasets are needed in a variety of application areas.

Long-term Vision

- Construct new benchmark datasets in a variety of application areas. These will provide methods of evaluation (e.g., IT Automation).
- Release open-source code and synthetic data on platforms such as Github, Zenodo, and HuggingFace to help the advancement of the academic community on agentic FM research.
- Extend systems-theoretic concepts from information theory, control theory, game theory, and statistical learning theory to understanding the fundamental limits of problem solving.
- Ability of AI agents to delegate to more skilled agents.
- Middleware & runtimes for scaling agentic systems.
- Identify communication and collaboration patterns that yield collective intelligence, perhaps drawing on ideas from business process modeling.

4.1.1 LLM as an Abstraction (LLMaaA)

We envision that LLMaaA will create a paradigm shift that will enable a new class of applications that can adapt to changing computational demands in real-time, efficiently use diverse computing resources and various agents (either LLM-based or non-LLM-based), and provide intuitive, natural language interfaces for complex tasks.

Challenges

- Provide intuitive, user-friendly, natural language-based, interface for building, deploying, and managing complex user applications.
- LLMs often struggle to adapt to task-specific requirements without human intervention.
- Lack of reconfigurability makes it difficult to adapt LLMs to different tasks without extensive engineering efforts.

- LLMs often fail to demonstrate evolvability, as they do not automatically learn from new data in an on-line fashion and may struggle to keep pace with rapidly evolving environments, requirements, and developments.
- LLMs may be plateauing in capabilities and there may be need to draw on hybrid approaches that use LLMs as subcomponents.

Long-term Vision

- LLMs become building blocks for establishing a brand-new abstraction equipped with user-friendly application interfaces, and larger hybrid compound AI systems.
- LLMaaA abstracts away the complexities of traditional programming and system management that are non-uniform and complex.
- Adaptable, evolvable and robust LLMs.
- Leverage advanced cloud application platforms that host LLM and non-LLM agents, provide user-friendly, natural language-based, secure, scalable, and evolvable services for building, deploying, and managing complex applications.
- Build and create all these afore-mentioned agentic features in a unifying framework (we call it THINKagents – details in the vision white paper), so it becomes the enabler of LLMaaA.

4.2 AI Model Optimization and Runtimes

Sub-area Leads: Sara Kokkila-Schumacher (IBM), Charith Mendis (Illinois), Mudhakar Srivatsa (IBM Advisor)

The rate and pace of generative AI innovations are staggering. New models and use cases are emerging every day, and potential applications abound. While the disruptive potential of these technologies has been established, the software stack for optimally serving them is not mature and hardware innovations to accelerate inferencing are rapidly evolving. To quickly get these models from discovery into productive use in the real world, a systematic approach to designing the software stack and hardware for inferencing is needed that applies across model architectures, across hardware platforms (CPU, GPU, and evolving accelerators such as the IBM AIU Spyre) and for small and large models alike. Today's AI model serving landscape is brittle. Finding the right software optimizations and runtimes can provide multiple orders of magnitude improvements in latency and throughput, but these optimizations are not universally applicable, and their adoption is difficult to automate.

Novel neural architecture designs and training methodologies will evolve as foundation models grow in popularity. Novel optimizations at the compiler and framework level are needed as AI becomes pivotal in enabling new frontiers in future scientific discovery.

We seek research proposals that will explore all aspects of this emerging area. Suggested high-level problem domains include the following.

AI Compilers and Runtimes

- Focus on AI compilers, framework-level optimizations, common accelerator abstractions and high-level kernels (e.g., OpenAI Triton), and leveraging hardware innovations to enable efficient scaling and optimization of emerging AI models, maximizing their impact on real-world applications.

- Scale FMs (Foundation Models) to handle larger context lengths is crucial for advancing AI applications in areas like NLP, climate prediction, and geospatial data-driven applications.
- Scale sparse models which play a pivotal role in predictive analytics for fields like drug and material discovery and quantum chemistry, but are challenging to scale due to irregular data.
- Advance the IBM AIU Spyre accelerator software stack.

End-to-end Edge-Cloud Transformation and Optimization (more details in Section 4.2.1)

- Optimize the balance between edge and cloud AI computation based on flexible SLOs, leveraging fine grained composable acceleration.
- Focus on unified programming models, specialized data communication methods, co-design of neural architectures and accelerators, and integrated offline and online optimization techniques.
- Extend end-to-end system prototypes and benchmarking that are essential to validate these ideas.

Challenges

- Handle long context lengths with >1 million tokens.
- System-level scaling by employing novel hybrid parallelism strategies to accommodate different components in the hybrid model.
- Agile and retargetable compiler construction for multiple accelerators, using formal methods or machine learning techniques.
- Unified programming model (e.g., OpenAI Triton) for heterogeneous hardware.
- Accelerating FM inference, especially in resource constrained environments, without sacrificing accuracy.
- Adapting FMs to new data domains with limited supervision.
- Exploiting full potential of FMs beyond fine-tuning.
- Self-directed enhancement of FMs for novel downstream tasks.
- Developing FMs with domain-specific knowledge.

Long-term Vision

- Novel model architectures such as SSMs, memory-augmented neural networks, linear attention mechanisms that reduce the computational cost of LLMs, while preserving accuracy.
- Novel optimizations both at the compiler and distributed framework level are needed to scale sparse ML models and enable the next generation AI applications that revolve around sparse data.
- The same programmability and compiler support that is available for more established hardware platforms such as GPUs and CPUs should become available to make emerging accelerators more mainstream, including IBM's AIU Spyre line of accelerators. To achieve that, innovations on programming languages, as well as novel compiler construction methodologies will be extremely important.

4.2.1 Enabling the Edge Transformation: End-to-end Edge-Cloud Deployment

The rapid advances in FMs, Gen-AI, and LLMs are fundamentally changing how we interact with computing at the edge, e.g., with immersive computing or extended reality (XR), robotics,

autonomous vehicles, drones, etc. These new modalities for computing have the potential to transform most human activities.

Challenges

- How to design accelerators for the diversity and evolution of emerging workloads, in an integrated edge-cloud ecosystem?
- How to design a uniform communication interface that serves the needs of diverse accelerators and cores in a system-on-chip (SoC), a system-in-package (SiP), or chiplets?
- Develop co-designed edge-cloud architecture.
- Specify flexible SLOs in multi-modal environments that reflect the user subjective view.
- Distribute computation between edge and cloud effectively.
- Disaggregated and composable fine-grained specialized acceleration.

Long-term Vision

- Can we have a uniform programing interface providing the ease of coherent shared memory but with the specializations?
- Co-design of neural architecture, accelerators, and inferencing workflows.
- Integrated offline and online model optimization techniques.
- Compiling and scheduling with flexible SLOs on flexible systems.

4.3 Hybrid Cloud Platform and Middleware

Sub-area Leads: Chen Wang (IBM), Volodymyr Kindratenko (Illinois), Carlos Costa (IBM Advisor)

Emerging compute-intensive workloads, ranging from state-of-the-art large-scale simulations to disruptive new AI/ML techniques, are becoming increasingly complex workflows that require diverse computational and data resources. These workflows encompass a wide range of computing and data needs, from the remarkable growth of large-scale distributed training leading to self-supervised models, also known as FMs (Foundation Models), to complex workflows that involve both asynchronous batch compute (such as data ingestion, pre-processing, simulation, and training) and synchronous interactive processes (such as model inference). They commonly span multiple stages that operate in different computing environments. Moreover, they benefit from and leverage an increasingly wide range of computing resources, including commodity CPUs, high-end GPUs, specialized AI accelerators, and upcoming quantum devices.

The proliferation of cloud computing technology has transformed how heterogenous computing resources can be used through multiple layers of abstractions, ranging from virtualization to container technology, as well as the extensive computing orchestration that enables elasticity, fault-tolerance, and flexibility through these abstractions. However, traditionally, the majority of distributed, large-scale applications and commonly used libraries for simulation have been written with programming models developed for a fixed, homogenous architecture in a confined environment such as MPI. While this fixed structure ensures low latency and efficient communication, it prevents this class of applications to benefit from cloud-native advantages such as elasticity, fault-tolerance, and portability, including hardware specialization abstraction.

With this RFP, we seek proposals that address the following dimensions:

Adaptive Middleware and Runtime in Hybrid Cloud Systems

- Develop adaptive and intelligent middleware and runtime solutions to optimize the interplay between compute and communication across distributed AI workloads.
- Design a unified control plane for AI-powered management of the resources in heterogeneous and dynamic cloud environments.
- The goal is to make future hybrid cloud dynamically adjust to real-time workload demands, evolving system topologies, and edge-cloud coordination, through AI-driven workload management.
- Advanced collective communication primitives, leveraging emerging technologies like UCIe, CXL, UALink, UltraEthernet, etc., in designing a uniform communication interface that addresses multiple granularity levels spanning from chiplets to intra-node and inter-node communication patterns.

Cross-Layer Automation and Integration

- Optimize cloud infrastructure for complex computations, cross-layer automation, integration, quality and cost control, and observability.
- Efficient resource allocation, scheduling, and SLO optimization, ensuring minimal delays and max availability.
- Develop automated frameworks for seamless orchestration and monitoring across cloud layers, which will improve efficiency, flexibility, robustness, adaptivity, and scalability.
- Robustness and System Health Monitoring for AI in Hybrid Cloud - Advanced fault detection, and self-healing algorithms and mechanisms to ensure long-term health, resiliency, and dependability.

Energy-Efficient AI Workloads and Sustainability in Hybrid Cloud Systems

- Allow AI runtimes to make energy-aware decisions during model training and inferencing, contributing to carbon efficiency while maintaining high performance.
- Develop AI-driven energy management techniques that dynamically optimize energy consumption across hybrid multi-cloud systems powered by diversified energy sources, including renewable sources.
- Adaptive techniques to balance energy and performance, efficient compute/memory/storage management, and smart workload distribution between edge devices and cloud resources.

Challenges

- How would the Cloud-Native runtime system support a myriad of middleware, ranging from classical HPC to Gen AI, and including big data processing and analysis, efficiently?
- Hardware Abstraction Layer (e.g., DRA) is needed to address the challenge of managing (e.g., in terms of control plane, communication, monitoring) hardware heterogeneity and specialized accelerators like GPUs, TPUs.
- Support for advanced programming models, including AI-powered workflow design tools.
- Adaptive Execution Engine to pave the way for workflow automation, lifecycle management, and the seamless handling and optimization of data.
- Integration with emerging technologies such as Quantum Computing.

Long-term Vision

- Unified Control Plane: used for AI-powered management of the resources in heterogeneous and dynamic cloud environments.
- High-Level Parallel Abstraction: investigate developing a programming interface that allows developers to express parallelism and data movement without delving into low-level details.
- Autonomous Runtime Optimization: needed to address the challenge of hardware specialization. As common patterns and libraries arise, an autotuning runtime should be employed to encode optimizations for more efficient computing, while keeping the underlying complexity hidden from users.
- Library of Optimized Computation Primitives: commonly used in AI/ML and simulation workloads. These primitives should be optimized for various hardware accelerators and accessible through the high-level programming interface. These libraries can then be utilized with standard APIs, irrespective of the offloading mechanism behind them.

Meanwhile, this sub-area also covers cross-layer optimization tasks. They are described in Section 4.3.1 and Section 4.3.2 below.

4.3.1 Robustness, Dependability, and Scalability

Challenges

- Correctness and fault-tolerance of scalable infrastructures and platforms are difficult. Without strong consistency for scalability, implementing correct, fault-tolerant controllers is challenging.
- Software dependencies and cross-system interactions are of immense complexity.
- New fault domains are exposed to cloud-based applications.
- Automated system operations become single points of failures.
- AI/ML robustness and interpretability are critically important when used for critical operations or system components.
- Scalability and security challenges still remain when various workloads of diversified types and sizes need to share the multi-cloud system efficiently in a dynamic fashion.

Long-term Vision

- Formally verified, provably correct cloud infrastructures and platforms.
- Cloud-native applications made reliable.
- Safe and reliable automated and AI-based operations.
- Agents and LLMs for incident management.

4.3.2 Energy Optimization and Sustainability

Challenges

- Server nodes in hybrid cloud are facing a significant challenge of heterogeneous hardware which needs optimized local control for energy efficiency.
- Hybrid multi-cloud systems are facing a significant challenge in terms of holistic resource management that needs to integrate workload scheduling, placement, server power state control, datacenter cooling system controls, and renewable versus non-renewable management.

- The challenge of using renewable energy for datacenters is their variability across time and space.
- Co-designing power and resilience management is required to provide fast failure recovery and differential treatment to critical/non-critical services to minimize disruptions while optimizing carbon footprint.

Long-term Vision

- Cooperative energy-optimization approaches between techniques of: (a) AI for Systems and (b) Systems for AI.
- Energy optimization and sustainability for AI workloads such as LLMs that exhibit exponential growth of data and model sizes and infrastructures used.
- Fine-grained energy / carbon measurement tools.
- Sustainable modular hardware and software: (a) high churn in the supply chain increases carbon footprint; (b) solve the mismatches between software and hardware lifecycles.
- Split reward models and failure recovery acceleration for SLO-aware energy optimization.
- Access to cloud testbeds with renewable energy and access to corresponding datasets/traces.

4.4 Infrastructure and Security

Sub-area Leads: Hubertus Franke (IBM), Nam Sung Kim (Illinois), Seetharami Seelam (IBM Advisor)

FMs (Foundation Models) represent an important paradigm-shift in AI. The use cases for FMs are growing by the day. Alongside the mounting excitement surrounding this next wave of AI is the realization that working with FMs is remarkably complex, and training them is computationally intensive. Training models with billions of parameters can easily consume hundreds to thousands of state-of-the-art GPUs, running continuously for weeks. In short, we need AI-optimized computing.

This RFP is interested in proposals seeking to explore innovations that could decrease systems cost and/or increase performance for large-scale distributed AI systems. Some example topics we would like to explore:

Unified, Programmable, and AI-Optimized Networking for Distributed AI Workloads

- Design unified, reconfigurable, and AI-optimized software-defined networking that facilitates secure and efficient data transfer across a shared hybrid cloud environment, eliminating inefficiencies and security vulnerabilities in inter- and intra-node communications.
- AI-driven network orchestration schemes that intelligently allocate bandwidth and resources, minimizing latency and cost while maximizing throughput for AI frameworks, like PyTorch, and ensuring secure data flow across diverse networking protocols (e.g., RoCE v2, InfiniBand).

Advancing AI Systems through Co-Design and Emerging Hardware Innovations

- Streamline data transfers and optimize memory access leveraging system co-design and emerging hardware technologies such as Compute Express Link (CXL), Advanced Matrix Extensions (AMX), and GPUDirect.

- Software-defined interfaces and cache-coherent interconnects need to be co-designed to improve fine-grained computational efficiency, system security and isolation across AI workloads.

Data Management and Storage Efficiency for Large Models

- Place and migrate data in real-time, ensuring efficient storage use while reducing bottlenecks – not just within a single cloud, but also across edge devices, private clouds, and public clouds.
- Innovate data management systems that intelligently distribute and manage large amounts of data (e.g., data used by FMs) across AI-Accelerator/CPU memory, SSDs, and node-local storage in hybrid cloud environments securely and efficiently.

Adaptive and Reconfigurable Cloud System for AI Workloads (more details in Section 4.4.1)

- Enhance reconfigurability and adaptability in hybrid cloud systems through technologies such as programmable SmartNiCs and in-network switches, reconfigurable hardware and interconnects, software-defined programmable interface, and workload-adaptive control strategies.
- Coordinated reconfiguration and specialization tailored towards specific workload characteristics will enable clouds to achieve significant performance gains, making them more affordable.

Challenges

- AI model architectures, and thus their resource utilization characteristics are changing rapidly as they try to address needs in multi-model domains, support large context windows, and support emerging agentic use cases.
- Systems technology such as accelerator, accelerator interconnects, networking, storage, protocols, software stack are changing at the same time. So, a primary challenge is to pick workload characteristics that are invariant and demonstrate new technology that addresses these workloads.
- Multi-cloud systems pose new security challenges to provide continuous attestation capabilities for heterogenous hardware and handling unreliable connections to remote systems.
- Security and confidentiality are more challenging with the growing complexity of the system.

Long-term Vision

- Develop system co-design ideas, to dramatically improve cost-performance, utilization, and energy efficiency of AI systems.
- Exploit new emerging hardware technologies such as CXL, AMX, GPUDirect, and various accelerators integrated with CPUs and NICs.
- Solve memory wall problem – large models are memory bound, limiting compute efficiency of accelerators. Emerging technologies like UALink, UltraEthernet, SmartNICs, acceleration on CPUs using technologies like AMX, emerging cache-coherent interconnects like CXL can be exploited to alleviate this problem.
- Easier and broader cooperative heterogeneous computing among the compute devices and network/storage/memory devices potentially with near-data processing (NDP) capabilities.

- Develop new security and confidential computing solutions to work with workloads of diversified types and sizes that share the multi-cloud system in a dynamic fashion, aiming for guaranteed security and SLOs.

4.4.1 Application-Adaptive Cloud System for Dynamic AI Workloads

Just like how an ASIC or an FPGA design can beat the general-purpose CPU in terms of performance and energy efficiency by a large margin (sometimes by 100-1000x), we envision that an application-adaptive cloud system would demonstrate a large advantage over the conventional general-purpose system as well.

Challenges

- Identify and overcome the limitations of current reconfiguration and programmability capabilities in cloud systems.
- How to achieve the desired balance between application specificity and flexibility? This will require innovations across multiple layers of the cloud stack.
- At the hardware level, how to ensure the seamless integration and co-design of technologies like CXL, programmable SmartNICs, and reconfigurable accelerators?
- Enable cross-layer programmability and reconfiguration across compute/storage/networking devices and platform/middleware frameworks will be essential to creating a system that can adapt in real-time to workload requirements.

Long-term Vision

- Develop novel reconfigurable compute, storage, and communication units for hybrid cloud.
- Incorporate predictive analytics for optimal workload and data placement and enhance the ability of platforms to operate across hybrid infrastructures.
- Explore how the system can adapt to varying task complexities, task arrival times, and hardware heterogeneity.
- Smart and distributed controllers with compute capabilities for connecting and managing different ranges of hardware resources dynamically to meet the changing workload demands.
- Coordinate different levels of reconfigurations in a coherent way for achieving higher performance gains.
- Adaptive and reconfigurable infrastructure for security.

5 Quantum Computing Thrust

Thrust Leads: Ruchi Pendse (IBM), Andre Schleife (Illinois)

In June 2023, scientists at IBM and UC Berkeley published their findings that the IBM Quantum Eagle system showed competitive performance to a supercomputer hosted by the Lawrence Berkeley National Lab in simulating spin systems evolving in time. The highly complex quantum circuits resulted in accurate answers when tested against a brute-force approximation algorithm. This result shows that present-day systems such as the IBM Quantum Eagle and Heron processors can provide value. Now is the time to explore the space of possible problems that can benefit from this value added by quantum computing. Please refer to the vision white paper introduced in Section 2 for more details.

5.1 The Path towards Quantum-Centric Supercomputing

Over the years, IBM Quantum has expanded its development roadmap to include software development that must go in conjunction with improvements in hardware. Present-day devices are greatly limited by a variety of errors that deleteriously affect the system.

Consequently, bearing in mind the advancements that have happened in the field since the inception of the Institute, the quantum thrust will be prioritizing the following areas:

1. Utility-scale applications and problems in the domains of quantum chemistry, material simulations, optimization, high energy physics simulations, etc. – running circuits with ~100 qubit circuit width and 1-5K gate depth.
2. Improvements in error mitigation schemes.
3. Algorithm development for expanding the quantum-utility application space.

5.2 Quantum-Utility Era Collaborations

Previous quantum systems were monolithic in nature. Scaling these systems required scaling all components simultaneously, which is, long term, not a feasible approach. Quantum-centric supercomputers (QCSCs) represent the next generation of scalable quantum systems, ones which will leverage best-in-class capabilities (both quantum and classical) to perform otherwise-intractable computations. QCSCs utilize a modular architecture to enable scaling and integrate quantum computation to increase the computational capacity of the system and use a hybrid cloud middleware to seamlessly integrate quantum and classical workflows. Developing, prototyping, and fielding such systems represent one of the most important tasks the quantum computing industry must do in the next decade. This will benefit from research in near-term systems that leverage both quantum and hybrid cloud systems.

6 Materials Discovery Thrust

Thrust Leads: Teo Laino (IBM), Huimin Zhao (Illinois)

At IBM Research, we see the future of materials discovery as dependent on a new paradigm where advanced AI methods, autonomous agents, and multimodal models come together seamlessly. As we move beyond traditional, manual workflows, the focus shifts toward AI-driven systems that are not only automated and parallel but also capable of learning, adapting, and integrating complex data streams. This evolution—driven by technologies like Hybrid Cloud and Quantum Computing—paves the way for self-sufficient agents and multimodal approaches that can accelerate breakthroughs in materials science in ways previously unimaginable.

Our focus for this RFP cycle is on the strategic enhancement of multimodal models, autonomous agents, and distributed model development methodologies, all centered around the augmentation of the Granite series of FMs (Foundation Models). As laboratory operations integrate more cloud-based functionalities, the application of these technologies is critical for creating agile, AI-driven, and scalable workflows in materials science. Please refer to the vision white paper introduced in Section 2 for more details.

Proposals should emphasize innovative solutions for foundational challenges in material design, using collaborative and integrated approaches with the Hybrid Cloud and/or Quantum

Computing Thrust teams. Priority will be given to projects that work closely and effectively with these teams to advance shared strategic goals, while also tailoring their findings to the development of next-generation models in materials science. Research directions of interest include the following:

1. **Multimodal Model Methodologies**
   We seek projects that push the boundaries of multimodal models for materials discovery by exploring early, late, and post-fusion techniques. Emphasis should be placed on:
   - Developing novel fusion architectures that integrate heterogeneous data types from a variety of sources (e.g., spectral, structural, computational data) to build more comprehensive material representations.
   - Investigating post-fusion strategies to enhance model interpretability, adaptability, and accuracy for real-world material science applications.
   - Implementing solutions that can scale across diverse material domains and data modalities, augmenting the capabilities of the Granite series of models.

2. **Autonomous Agents for Advanced Computational Tasks**
   Proposals in this area should focus on developing methods that allow AI agents to autonomously identify, formulate, and execute tasks in computational environments beyond the native capabilities of language models. Core areas of interest include:
   - Implementing self-learning agents capable of integrating computational tools into workflows autonomously, adapting to complex challenges in material science.
   - Methods to extend the capabilities of Granite models in achieving autonomy in computational tasks relevant to material design and discovery.

3. **Multimodal Synthetic Data Generation**
   Proposals should address the development of methods within the context of InstructLab to generate high-quality synthetic data across multiple modalities for use in training and validating models in material science. Emphasis should be placed on:
   - Development of technologies for creating robust, synthetic datasets that mimic real-world variability across different material types, properties, and conditions, enabling models to learn from enriched data representations.
   - Techniques to combine and harmonize synthetic data across modalities such as spectral, structural, and compositional data for improved model accuracy and generalization.
   - Leveraging synthetic data generation to support training of Granite models, enhancing their adaptability and performance in new or data-scarce material design and discovery tasks.

A fundamental requirement of this RFP is that all proposed methods be grounded in and contribute to the capabilities of the Granite series of models. Development contributions should be focused on (1) Extending Granite models' applicability across diverse material design and discovery tasks; (2) Building new functionalities or augmentations within Granite that enhance its utility for multimodal fusion, autonomous agent-based learning, or distributed model training; (3) Ensuring that all methodologies developed are adaptable, transferable, and open to iterative improvements that align with Granite's evolving framework.

Each proposal should:

- Clearly outline how it addresses one or more of the focus areas.
- Specify how the research will augment the capabilities of Granite models.

- Detail a collaborative plan that supports cross-institutional or cross-disciplinary efforts, if applicable.
- Include milestones, deliverables, and a timeline for integration of results with Granite model applications in material design.

7 Climate & Sustainability Thrust

Thrust Leads: Hendrik Hamann (IBM), Jingrui He (Illinois)

Climate and sustainability represent a very broad topic that cuts across many thrusts within the Institute, which aligns with the IBM's priority areas. In particular, the 2023 IBM Impact Report highlighted a number of IBM's sustainability initiatives, solutions, and applications of IBM technology. In this thrust, we are particularly interested in innovative solutions that enable timely and accurate analysis and prediction of the changing climate and its impacts, supported by high-performance compute and innovations in cloud computing. Please refer to the vision white paper introduced in Section 2 for more details.

Among others, Geospatial FMs (Foundation Models) represent a key focus area for IBM Research. For example, IBM and NASA developed the first geospatial FM, which was pretrained on satellite (Harmonized Landsat and Sentinel-2, HLS) data. More recently, IBM and NASA released the open-source FM Prithvi WxC for a variety of weather and climate-related applications. Trained on 40 years of historical re-analysis weather data, the model can be readily finetuned for different use cases. In the future, we aim to pretrain FMs on observations versus re-analysis. Observation-to-Observation FMs will become "digital twins" suitable for accelerating scientific discovery.

In this thrust, building upon previous efforts from the Institute on multi-modal Geospatial FMs for climate and sustainability, we seek research proposals that continue to explore scalable and efficient architectures and approaches for developing FMs, which can deal with the multi-modal and high dimensional data (as required for climate and sustainability applications). This includes model size, FM composability, loss functions, and efficient strategies for pretraining and content generation. The goal of the research is to advance applications of finetuned FMs in the areas of climate and sustainability with special emphasis on scientific discovery. The application areas include climate impacts (physical, economic, etc.), climate mitigation (carbon sequestration, etc.), and greenhouse gas emission quantification. FMs for managing, controlling and planning the electric grid are of interest as well.

The core areas of focus include:

1. Architectures for FMs for Climate and Sustainability:
   - Multi-modal FMs (geospatial, vector, raster, text, time series, etc.)
   - Efficient architectures for pre-training FMs
   - Composability of FMs
   - Novel pretraining strategies
   - Large-scale graph networks

2. Finetuning to enable C&S Applications
   - Spatial, temporal and channel reconstruction tasks
     - Cloud removal

    - Forecasting
    - Modality blending (e.g., between different satellites)
    - Downscaling
- Applications in agriculture (land-use, climate impact, food security)
- Applications in power systems (power flow, optimal power flow, state estimations, contingency analysis)
- Other applications areas including nature-based carbon sequestration, GHG monitoring (CO2, Methane, etc.), extreme weather, etc.

Finally, the teams are encouraged to consider open-sourcing components of their work as to encourage the global open science community working on developing climate solutions, to leverage the latest tools and capabilities in the field.

8 Summary

This document is prepared by the technical leadership of the IIDAI Institute from IBM and the University of Illinois to identify the priority focus areas in the technical thrusts of the Institute, and in which project proposals are being solicited for funding. Through the launch of a new research funding cycle, IIDAI is determined to tackle the challenges facing current hybrid cloud systems and change their trajectory to become more affordable, programmable, adaptive, resilient, and accessible, empowering the next generation of AI-centric applications for novel scientific discoveries.