Semiconductor Research Corporation

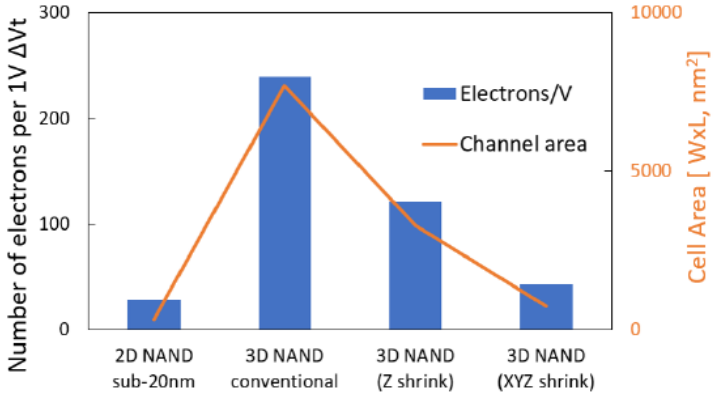
PRISM

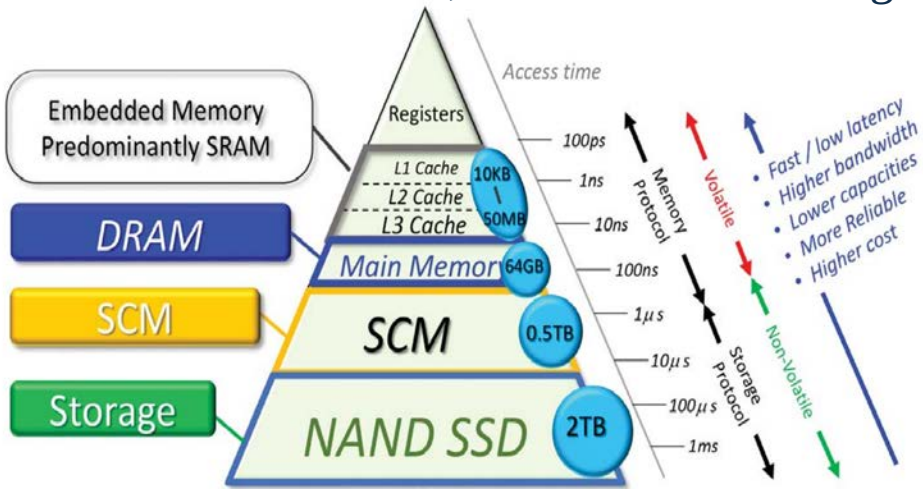
JUMP 2.0 PRISM Center Overview and Plan of Action

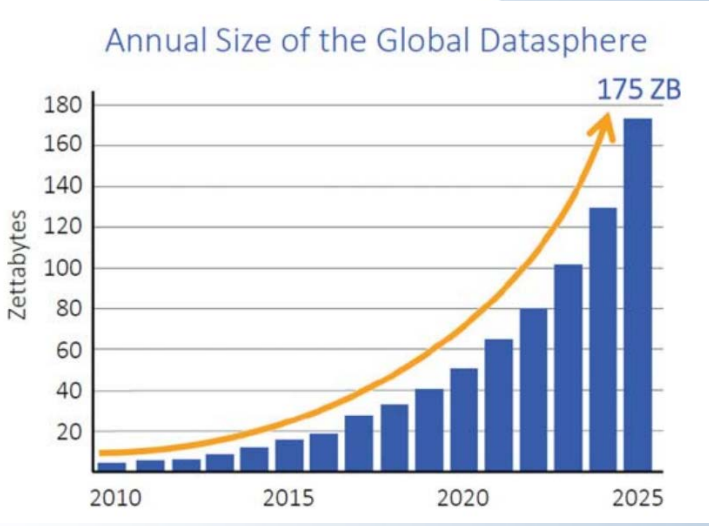
Science Advisory Board/Center Leadership Meeting
4/26/2023

Challenges

- Amount of data to be stored, moved & processed is rising exponentially
  - Global demand for memory/storage is growing rapidly, outpacing silicon production
  - Data movement is expensive
- Rising complexity makes programming and optimization harder
  - Heterogeneity of components and how they are integrated into systems
- Fundamental barriers to memory and storage technology scaling
  - Lower NAND string current, higher cell-to-cell interference, fewer electrons per stored state
  - Wordline disturbance, variable retenton time, reduced sense margin



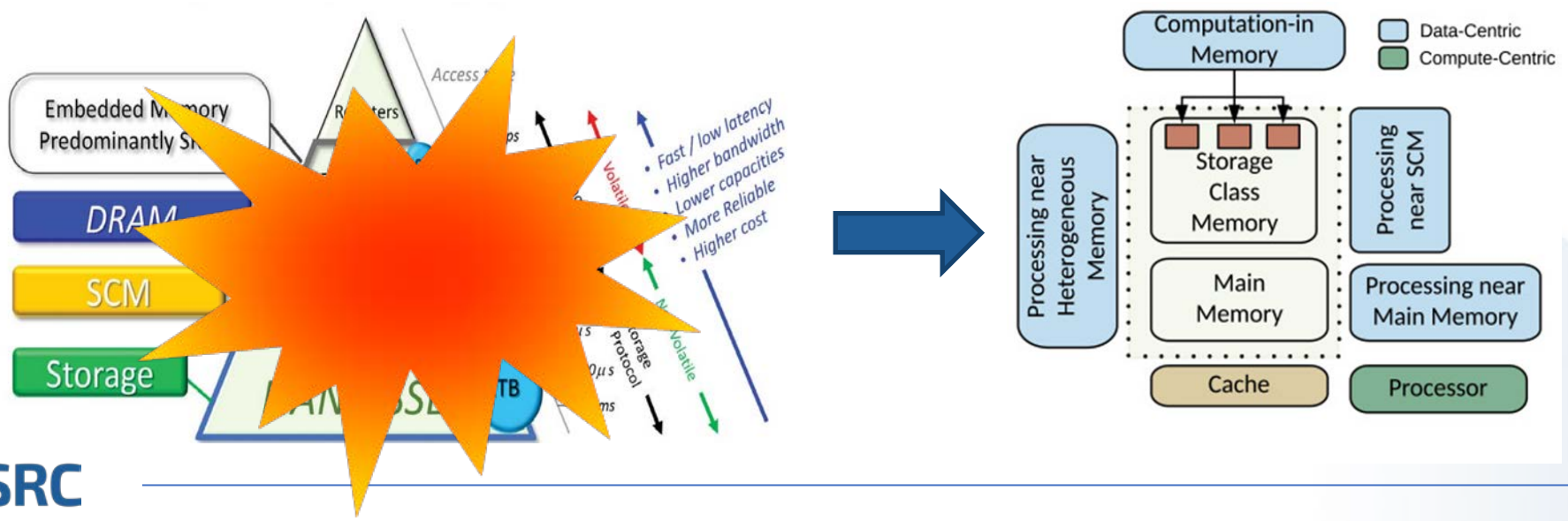




PRISM Vision

Solve fundamental intelligent memory/storage (IMS) scaling challenges for 2030

- Create a novel computing architecture that:
  - Answers when, where and how to store and process which data
  - Seamlessly integrates diversity of memory, storage, compute & software
  - Holistic cross layer IMS optimization from devices to applications
- Demonstrate capabilities using grand challenge applications



PRISM Overview

AIMS: Acceleration In Memory/Storage

Theme 4: Grand Challenges

Leads: Vijay Narayanan
Yizhou Sun

Personalized & Secure Drug Discovery

Deep Insights

Theme 1: Systems & Software

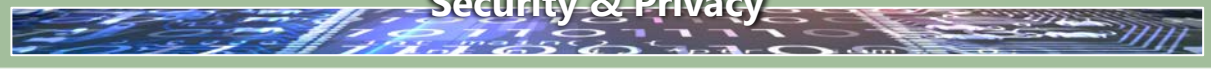
Leads: Emmett Witchel & Ada Gavrilovska

Platform Abstractions

AIMS Controls

Scheduling & Placement

Programming & Compilers

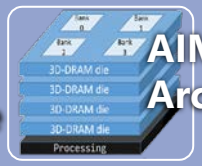


Security & Privacy

Theme 2: Architecture

Leads: Nam Sung Kim & Jishen Zhao

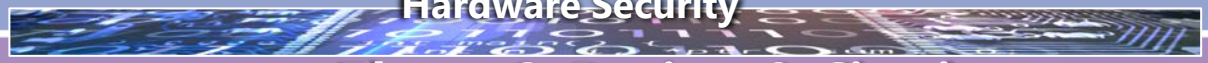
Memory & Storage Architecture



AIMS Architecture

Controllers & Interfaces

Disaggregated IMS Design



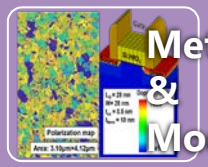
Hardware Security

Theme 3: Devices & Circuits

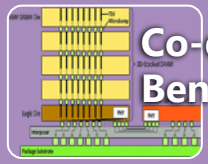
Leads: Suman Datta & Shimeng Yu

Intelligent Memory

Intelligent Storage



Metrology & Modeling



Co-design & Benchmarks

Emerging Memory

Cross-cuts

Performance
Programmability
Energy Efficiency
Security
Scalability
Virtualization
Composability
Reliability
Resilience
Availability

Baris Kasikci - Theme 1

Priyanka Raina - Theme 2

H.-S. Philip Wong - Theme 3

Eric Pop - Theme 3

Fredrik Kjolstad - Theme 1

Patrick McDaniel- Theme 1

Michael Swift - Theme 2

Nam Sung Kim - Theme 2

Vikram S. Adve - Theme 1

Jason Cong - Theme 2

Yizhou Sun - Theme 1

Sayeef Salahuddin - Theme 3

Sang-Woo Jun - Theme 2

Franz Franchetti - Theme 1

Vijaykrishnan Narayanan - Theme 2

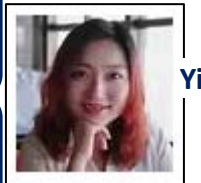
Kevin Skadron - Theme 2

Ada Gavrilovska - Theme 1

Shimeng Yu - Theme 3

Suman Datta - Theme 3

Tajana Simunic Rosing - Theme 2



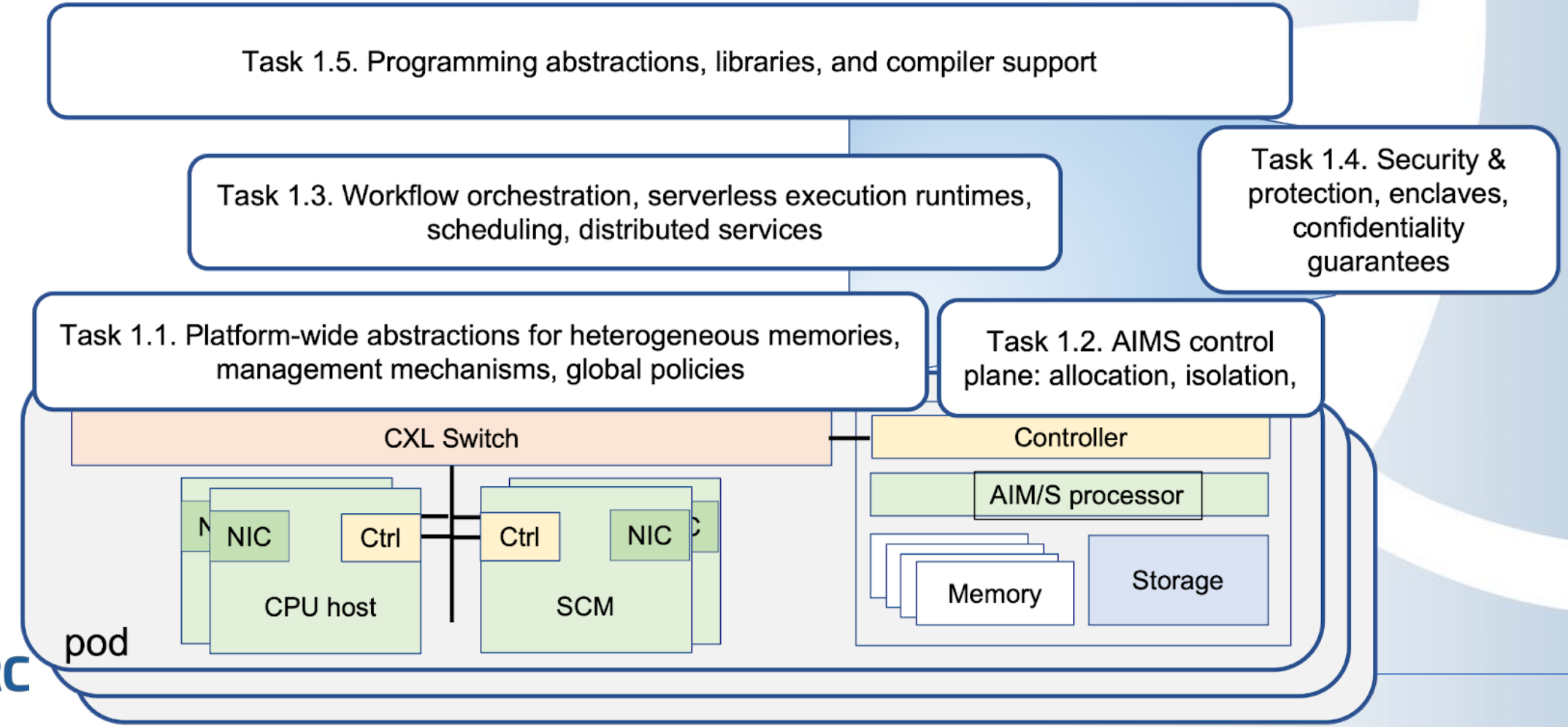
Yiying Zhang - Theme 1

Jishen Zhao - Theme 2

Emmett Witchel - Theme 1

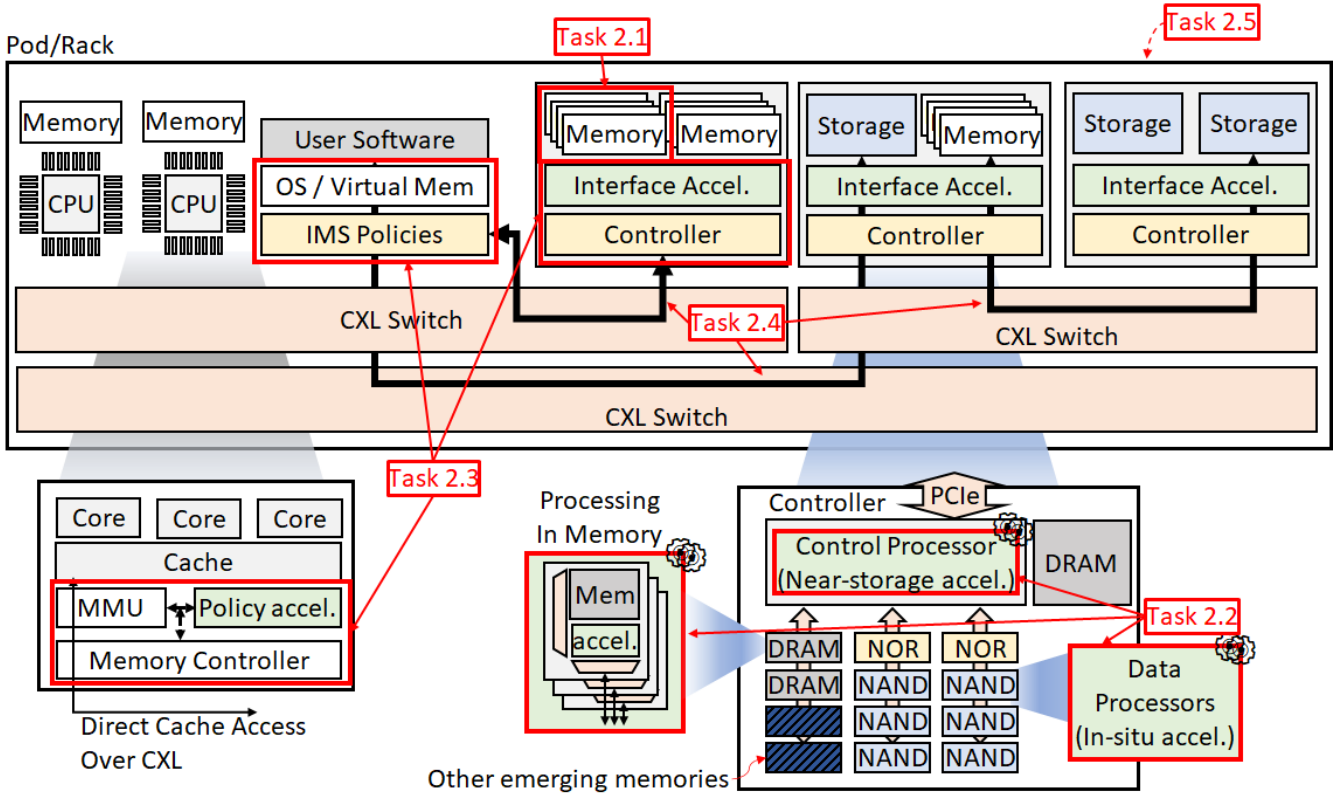
Theme 1: Systems and Software

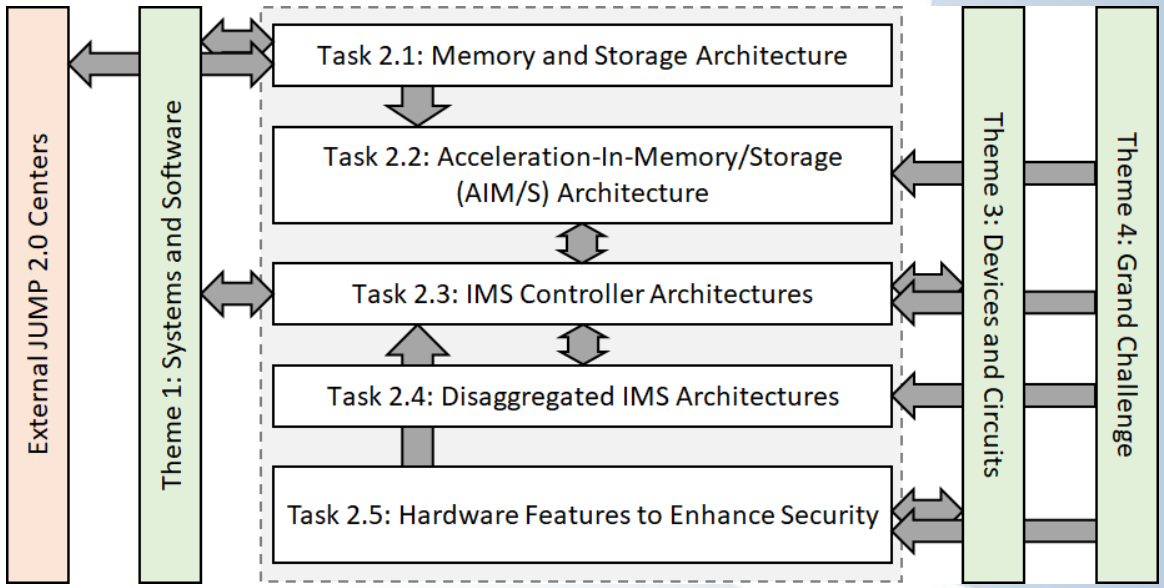
Goal: seamless deployment of grand challenge applications in virtualized & distributed IMS systems with 100x improvement



Theme 2: Architecture

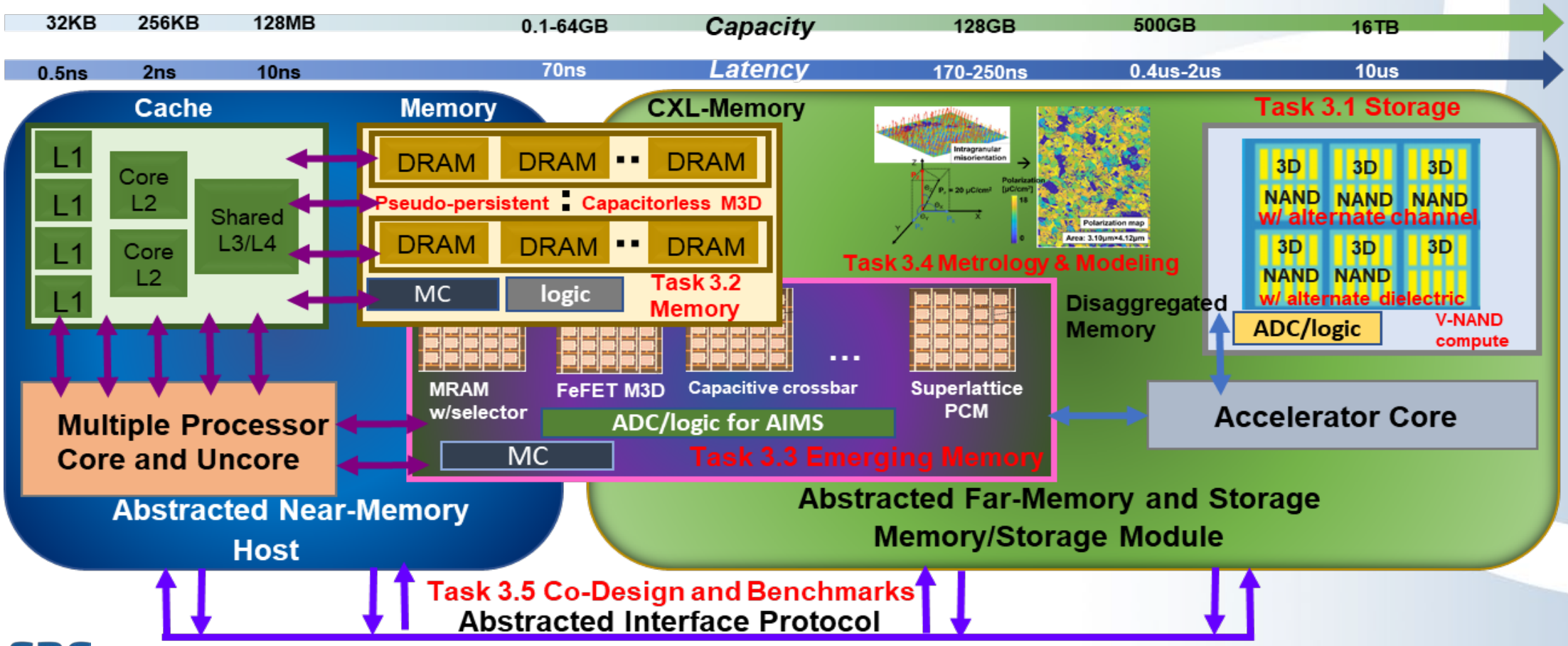
Goal: memory/storage architecture enabling 100x more powerful IMS computing capability at 10x larger capacity





Theme 3: Devices & Circuits

Goal: fundamental advances in devices and exposure of their controls to higher layers leading to 100x improvement in PPAC



Theme 4: Grand Challenge Applications

Task 4.1: Personalized & Secure Drug Discovery

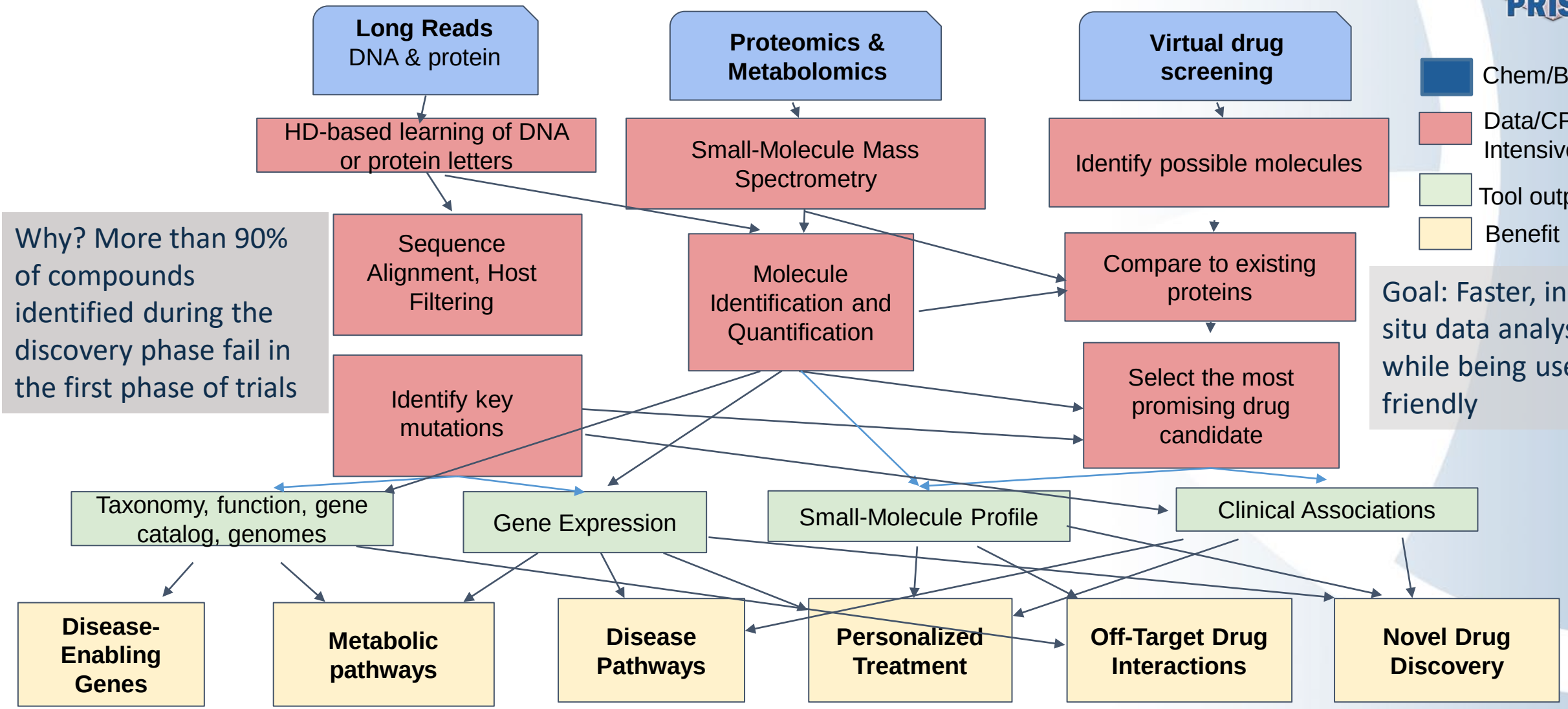
Task 4.2: Deep Insights Pipelines

Application Constraints & Needs
Code Repositories
Application Characterization
Algorithmic Innovations

Software Repositories,
Simulation tools,
Emulation Systems,
Chip tape-outs,
End-to-End prototypes incorporating PRISM ideas

PRISM Themes 1, 2 & 3

Personalized new drugs



Why? More than 90% of compounds identified during the discovery phase fail in the first phase of trials

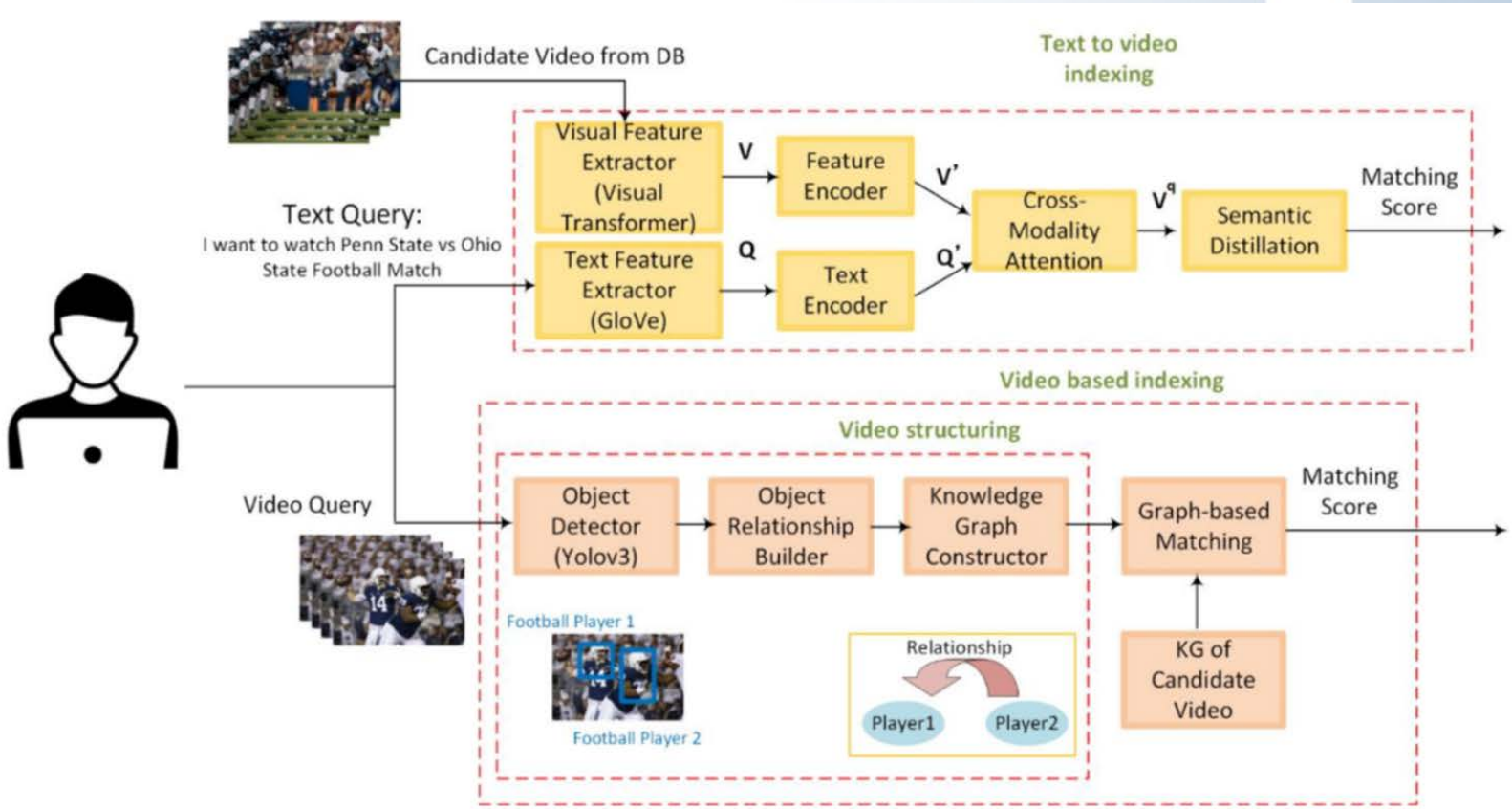
Goal: Faster, in-situ data analysis while being user friendly

Deep Insights

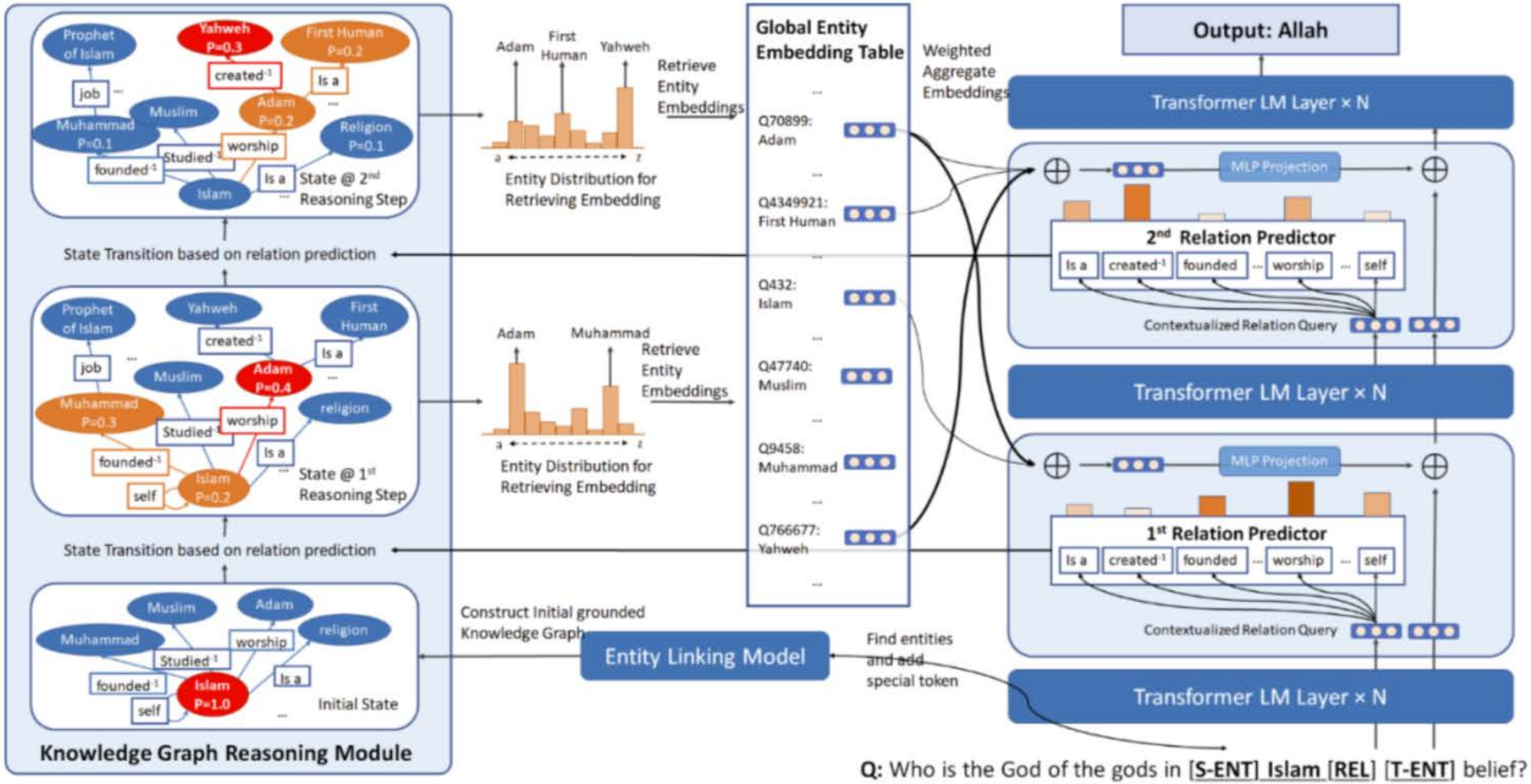
- Motivation: Most data generated will never be touched by humans directly and need automated analysis or human-machine partnerships
- Goal: 100x gain in power-performance-area-cost (PPAC)
- Key focus:
  - Deep reasoning to support open ended queries
  - Multimodal and cross-modal analysis
  - Distributed compute, memory and storage
  - Four end-to-end applications and datasets

Contextual Video Extraction with Text or Video Queries.

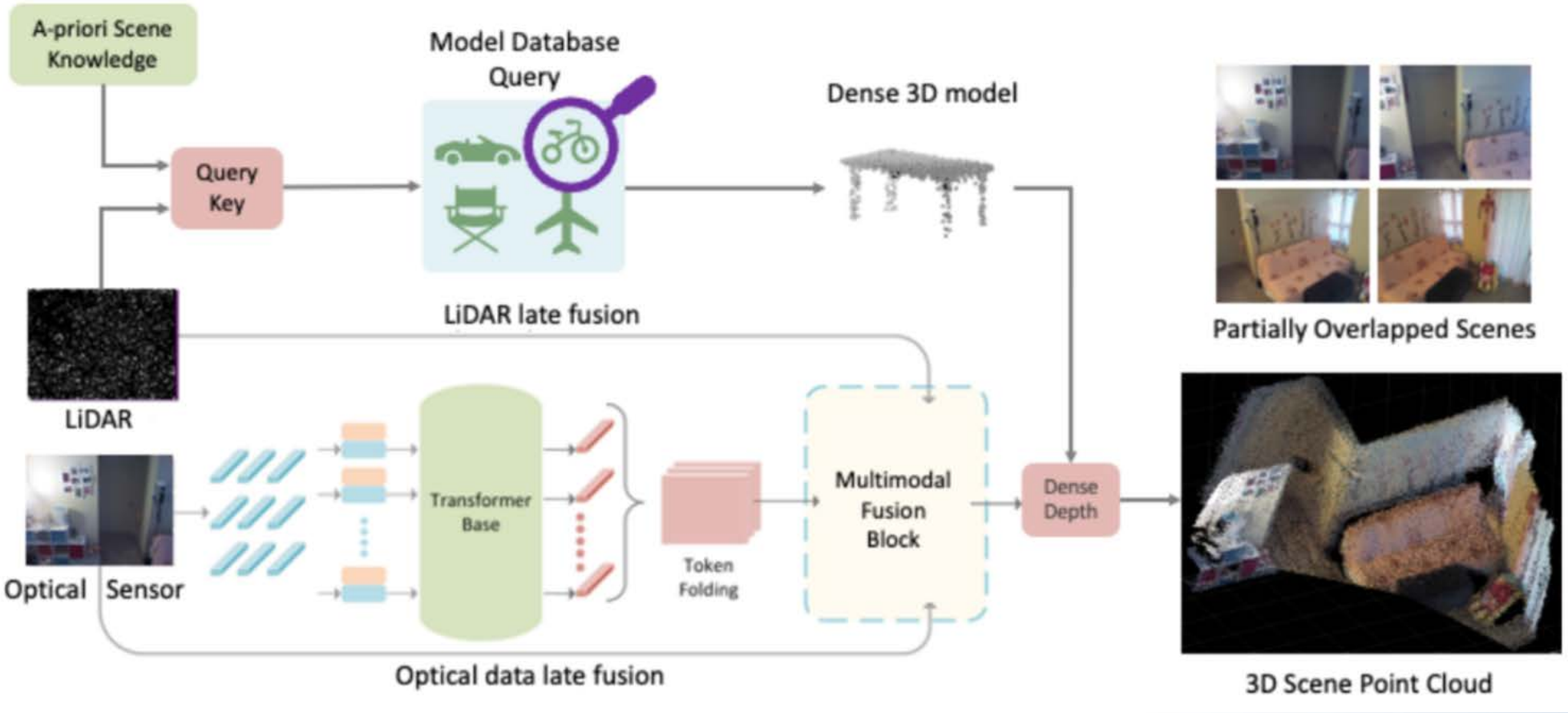
- >500 hrs of video/min uploaded to YouTube
- Video segments need to be retrieved using contextual information
  - Textual description of the scene(s) to be retrieved from the larger video database
  - A representative video clip input that queries other similar scenes in the video.



Integrating Reasoning Module into Transformer for Q&A



Contextual Fusion for Fast 3D Cloud Point Construction



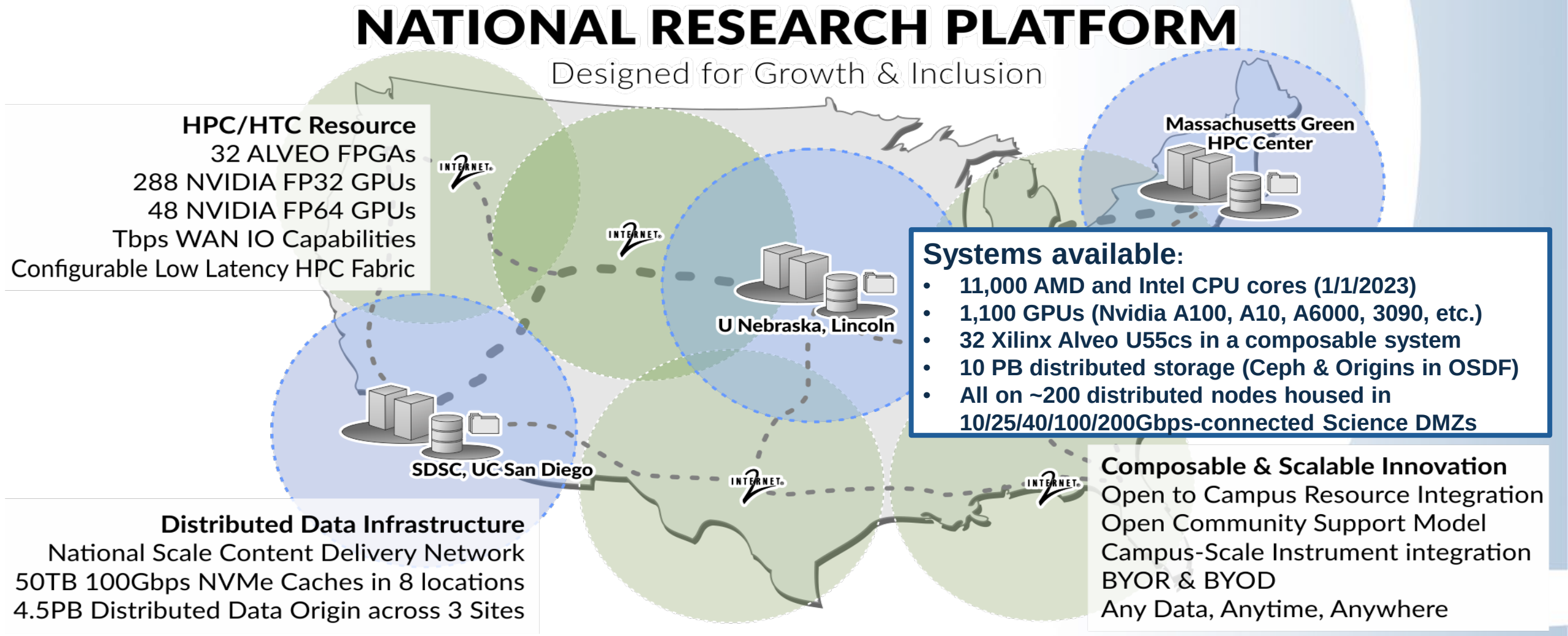
Collaboration with other JUMP 2.0 centers

- Theme 1 Cognition - CoCoSys: Tajana Rosing & Patrick McDaniel
  - HD computing for intrusion detection in hosts and networks
- Theme 4 Distributed Sys & Arch – ACE:
  - Jishen Zhao: Tiering of Serverless Snapshots for Memory-Efficient Serverless Computing
  - Ada Gavrilovska: End-to-end In-Fabric Programming for Graph Analytics
  - Kevin Skadron: Integrating Heterogeneous PIM in Disaggregated Systems
- Theme 6 Integration – CHIMES:
  - Nam Sung Kim: Reconstituted Wafer-based Heterogeneous Integration Technology Tailored for Memory and Storage Devices
  - Shimeng Yu: 3D NAND acceleration of Mass Spectrometry
- Theme 7 Devices SUPREME: Vijay Narayanan
  - Embedding Security into FeFET NAND Array Leveraging the Intrinsic Memory

Collaboration between PRISM themes

- Collaboration is under way; a few examples:
  - Systems to architecture:
    - Baris & Nam Sung Kim; Fred & Priyanka; Vikram &Tajana; Franz & Tajana
  - Devices to architecture:
    - Priyanka & Philip; Shimeng, Philip & Tajana
- Grand challenges theme:
  - Applications released to PRISM PIs and students
    - E.g. mass spectrometry CPU & GPU code plus datasets
  - Collaboration with
    - UCSD School of Medicine: Rob Knight (invited speaker today), Pieter Dorrestein
    - LLNL: drug discovery pipeline and molecular dynamics simulations – joint paper
    - LBNL: intrusion detection on networks; ESnet testing; student internship
    - DoD: deep insights theme
- Infrastructure for PRISM demos at UCSD introduced
  - PIs/students started using it, industry liaisons can join
  - Visit breakout session on Demo infrastructure today!

UCSD PRISM Infrastructure: Composable FPGAs, GPUs, CPUs



Visit NationalResearchPlatform.org to join


ACADEMIA • INDUSTRY • GOVERNMENT