HPC/AI Convergence

Mohamed WAHIB (RIKEN), Emmanuel Jeannot (INRIA), Artur Lorenzon (UFRGS), Philippe Navaux (UFRGS), Brian Spears (LLNL)

Central Point 1

Money will go after AI

Market will MOSTLY spec for AI

HPC has no option but to "shop" in that market

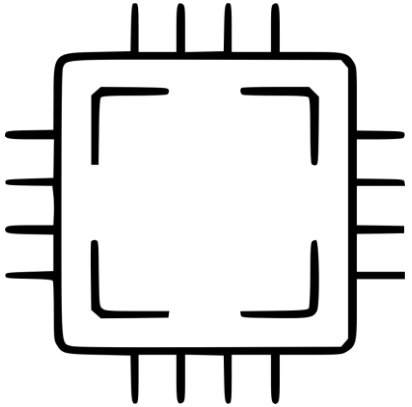
How much divergence between HPC and AI?

Central Point 2

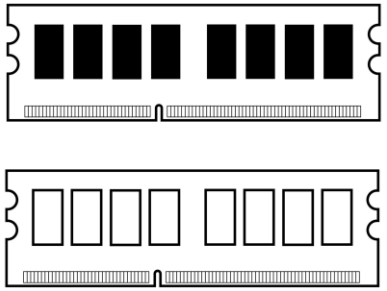
Will Supercomputers viewed as a service platform, and NOT a scientific instrument?

AI and HPC: Divergence vs. Convergence

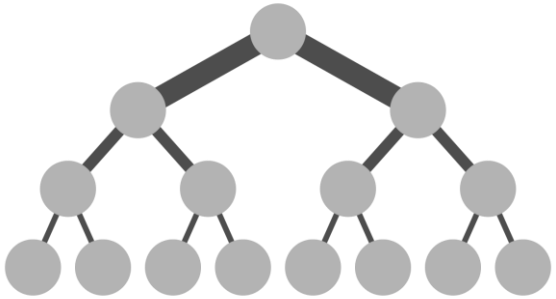
Compute



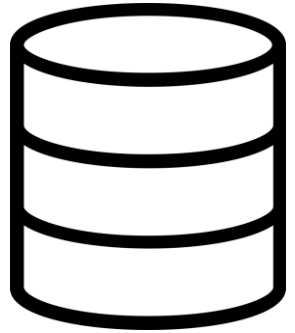
Memory



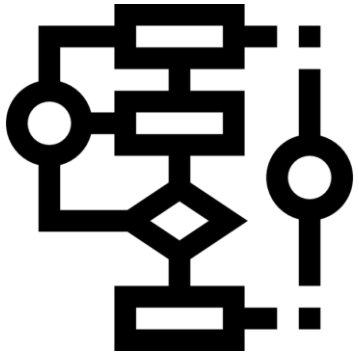
Network



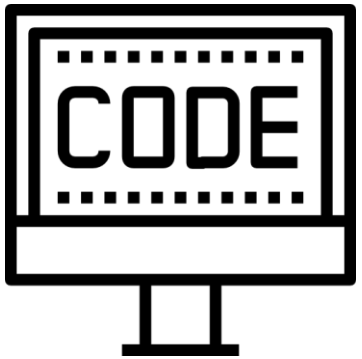
Storage



Algorithms



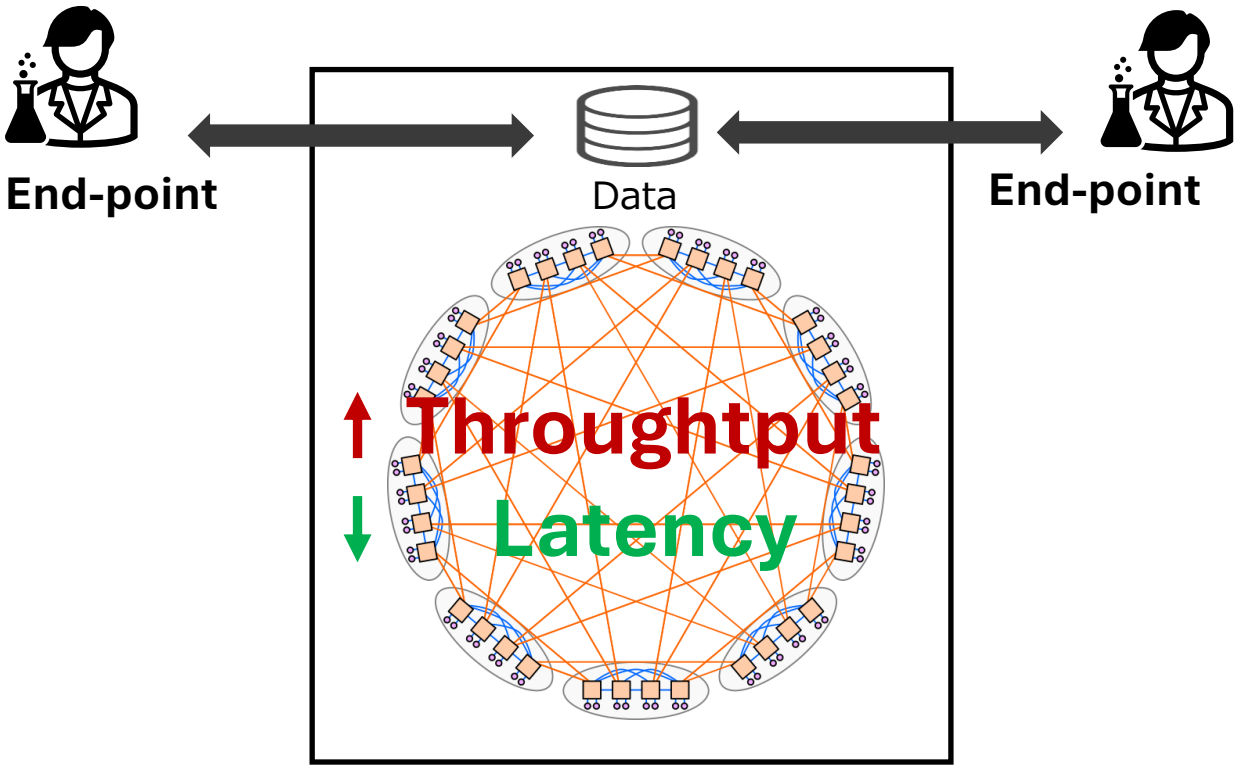
Programming



Libraries (Sys SW)

Pre-AI

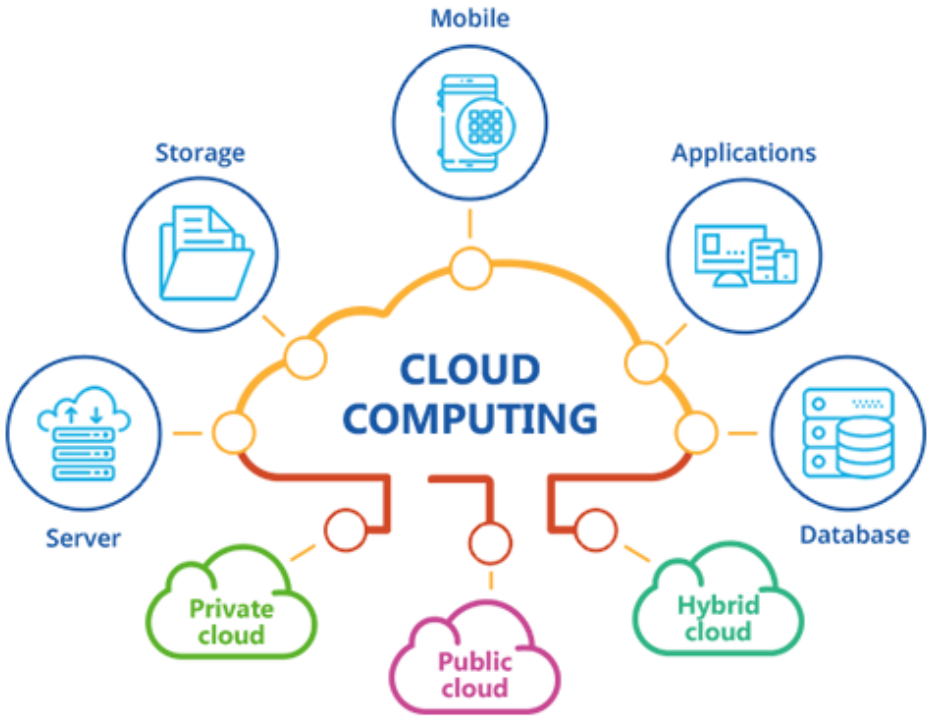
HPC Center



Design Philosophy:

- Scale-up + Scale-out (inside center)
- End-points to transfer data in/out
- 24/7/365 up time
- Minimum data pipeline software
- Supercomputer is an instrument, not a service

Cloud



Design Philosophy:

- Modest network (inside data center)
- Seemless integration to outside
- Redundency
- Rich ecosystem of data pipeline software
- Cloud is a service

AI Era

HPC Center

ModSim

Data

Scientific Instrument

Autonomous Lab

Memory

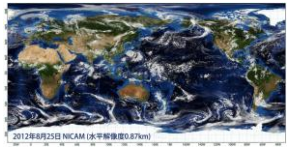
AI Model

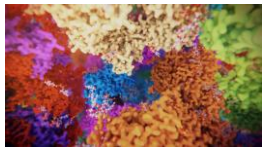
Access Protocol

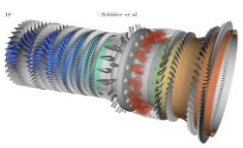
Throughtput

Latency

Physcial | Life/Medical | Engineering







LLM

Advanced reasoning AI model driving , discovering, and interpreting

Design Philosophy:

- **Scale-up + Scale-out (inside center)**
- **Seemless integration to the outside (agentic-style)**
- **Rich ecosystem of data pipeline software**
- **Supercomputer a service**

Evolution to AI-HPC Coupling

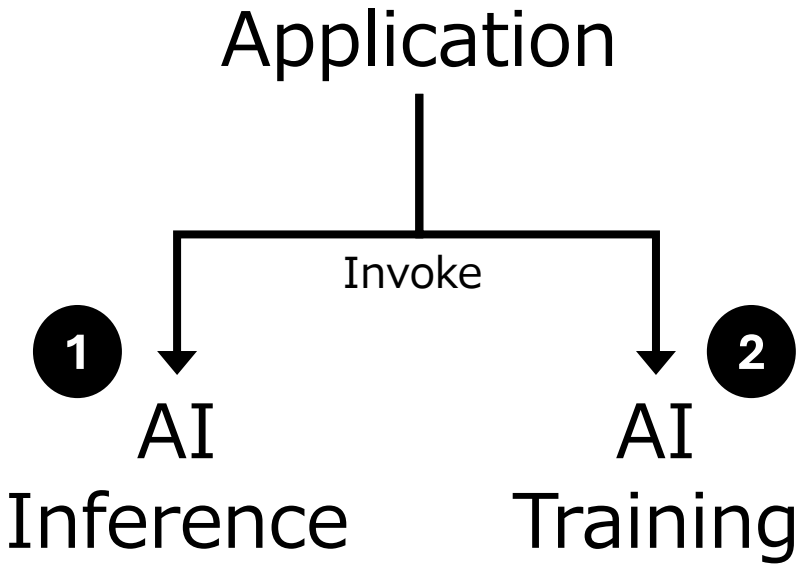
When coupling AI-HPC (vs. AI-only or HPC-only):

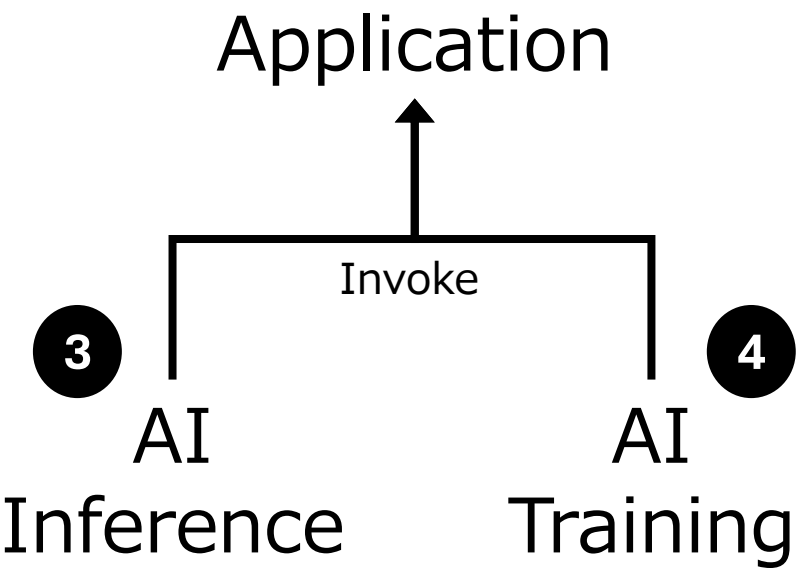
- AI-HPC (Application) Coupling Scenarios
  - Spec changes when coupling (vs. AI only or HPC only) Co-design
  - Software Stack Needed Sys Sw & Application
  - Resource Management Sys Sw & Application

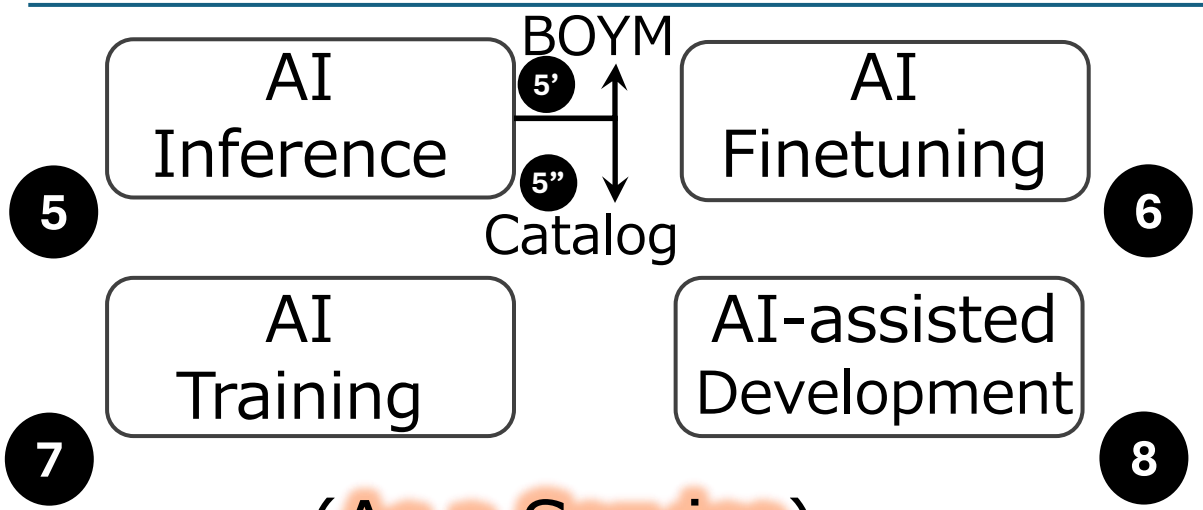
AI-HPC (Application) Coupling Scenarios

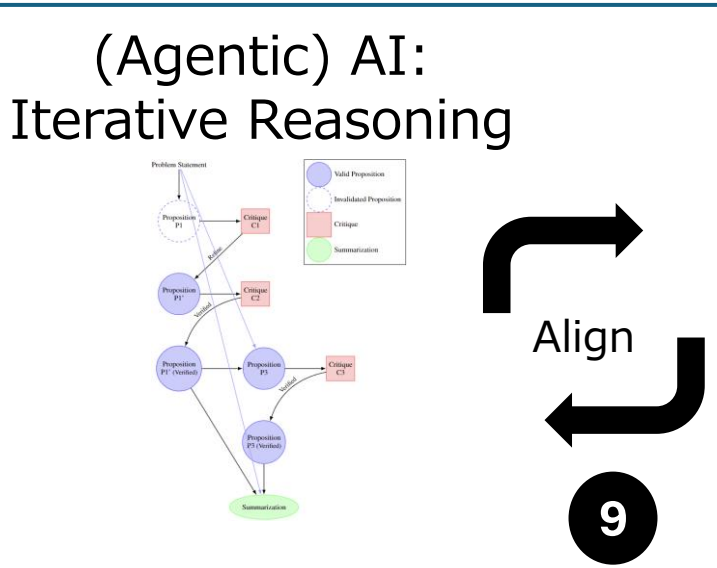
(AI-HPC Coupling ≠ Training/inference AI Models on HPC Systems)

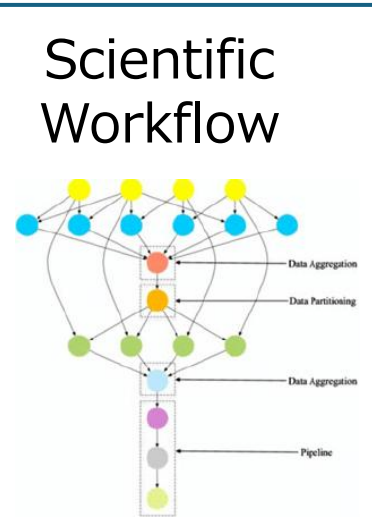
Enduser POV: Various Scenarios for AI-HPC Coupling



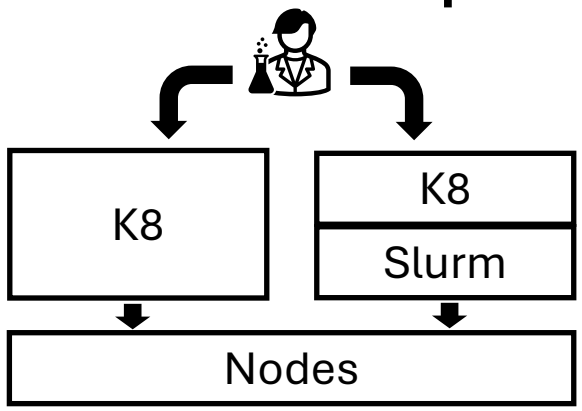




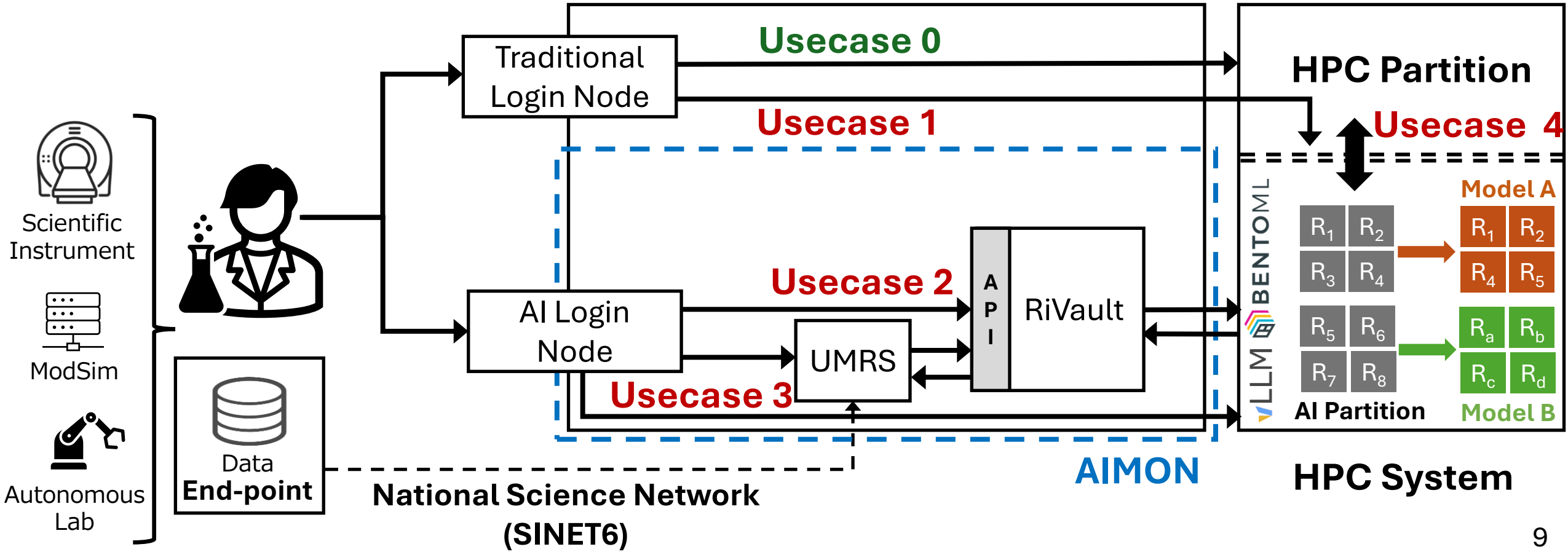




Supercomputers in the Era of AI + HPC Coupling



Usecase 0 = Traditional HPC
Usecase 1 = AI-assisted Coding
Usecase 2 = (LLM) Model Serving
Usecase 3 = Data + Ambient Model
Usecase 4 = Coupling AI & HPC



Spec changes when coupling (vs. AI only or HPC only)

Co-design

Questions to Answer

- Is interconnect (scale-up & scale-out), and storage
  - enough when AI-HPC coupling happens?
- How is AI and HPC workload co-located?
  - What level of co-location expected?
    - Break even point for co-location?
    - Scaling-up: split nodes to AI and HPC
    - Small scale: better to co-located (beware contention)
- Would coupling push to
  - heterogeneous resources?
  - heterogeneous deployment?
  - balance in precision in AI-HPC coupling?