Adding *Machine Learning* to the Management of *Heterogeneous Resources*

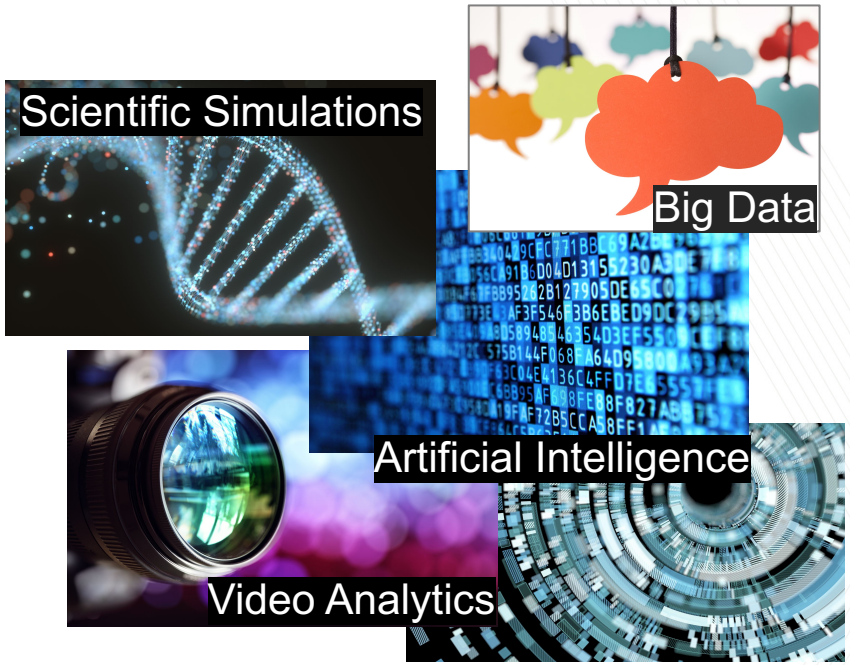
Thaleia Dimitra Doudali

The Era of Data

"More than **59 ZB** of data will be created, captured, copied, and consumed in the world this year."

Source: International Data Corporation, May 2020.

Exploded Data Sizes



Data Analytics Pipeline

ZBs of data

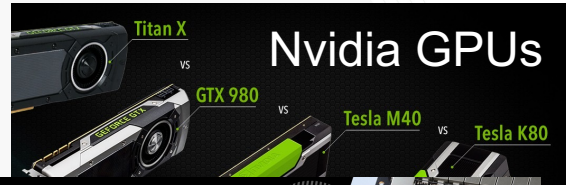
Need for speed and massive storage capacities!

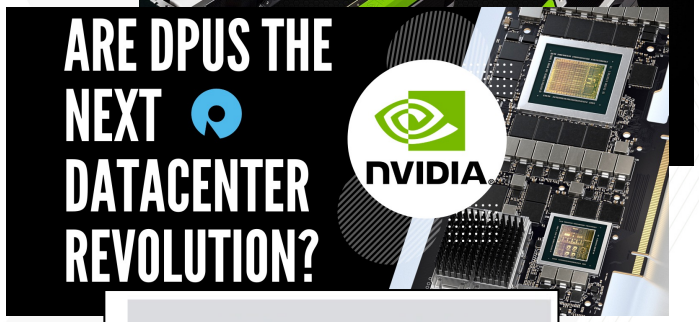
Georgia Tech
CREATING THE NEXT

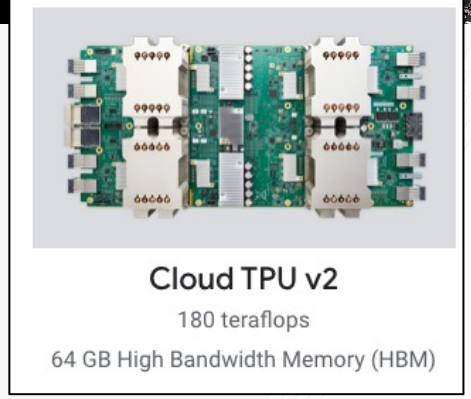
The Era of Heterogeneous Hardware

Emerging technologies across layers and vendors.

Compute Acceleration

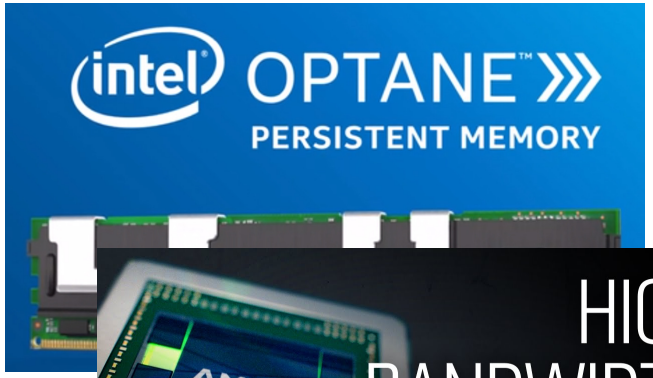


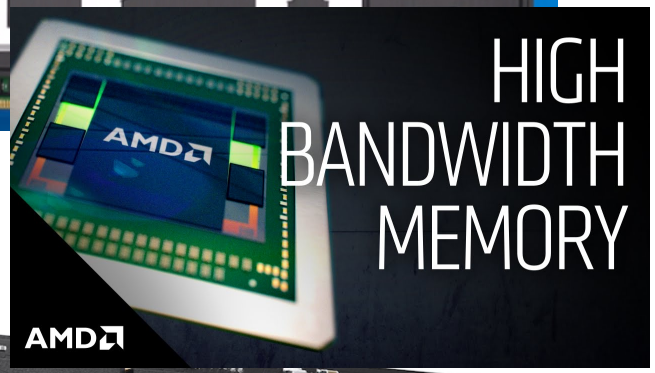




Google

Data Storage Acceleration





Network Acceleration

Mellanox Innova™-2 Flex
Open Programmable SmartNIC

Interconnection Standards

CXL Compute Express Link™

Georgia Tech
CREATING THE NEXT

The Era of Heterogeneous Hardware

Across computing platforms.

Supercomputers

HPCwire

Since 1987 - Covering the Fastest Computers in the World and the People Who Run Them

- Home
- Technologies
- Sectors

Sweden's National Supercomputer Center Launches 300 AI Petaflops System – 'Berzelius'

By Staff Reports

Number of Nodes	4,608
Node performance	42 TF
	512 GB DDR4 + 96 GB HBM2
	1600 GB
	>10 PB DDR4 + HBM2 + Non-volatile
	2 IBM POWER9™ 9,216 CPUs 6 NVIDIA Volta™ 27,648 GPUs
	250 PB, 2.5 TB/s, GPFS™
	13 MW
	Mellanox EDR 100G InfiniBand
	Red Hat Enterprise Linux (RHEL) version 7.4

Datacenters

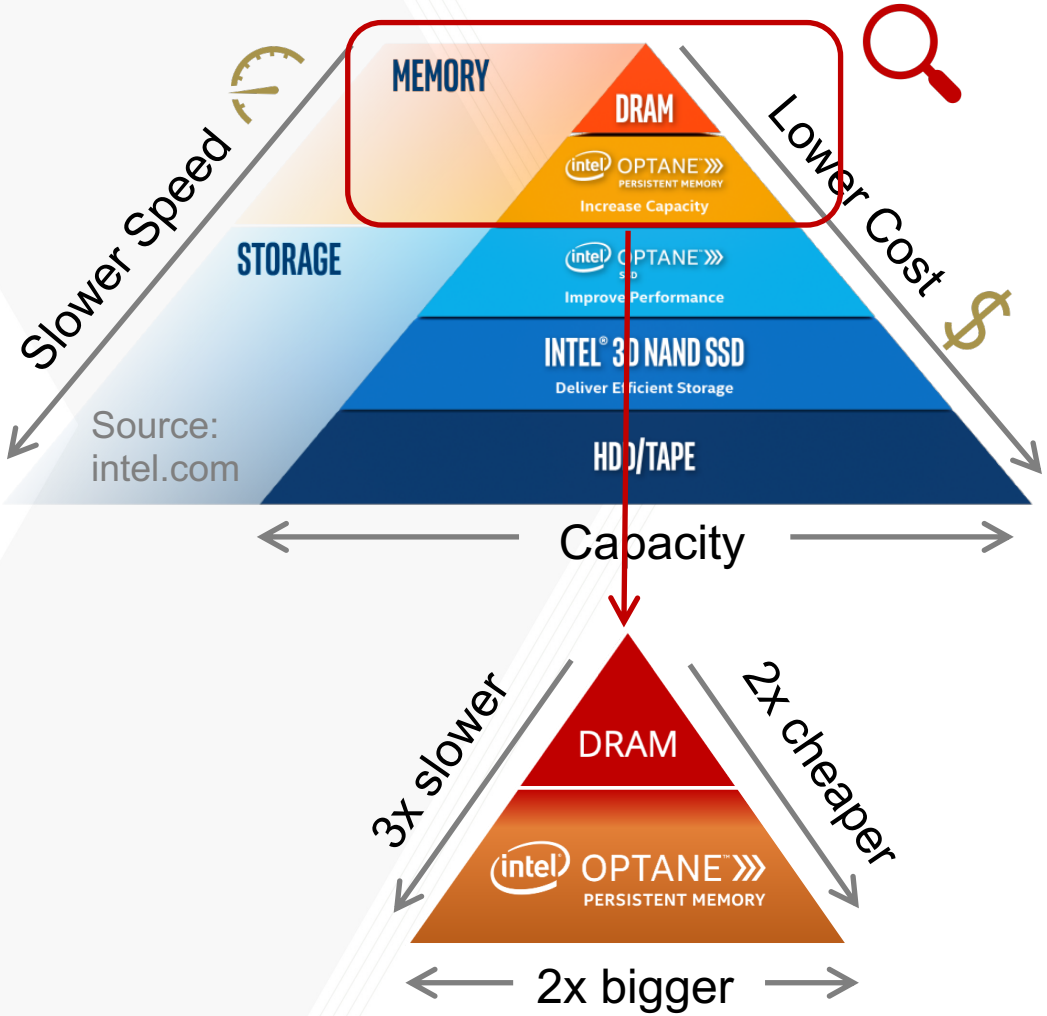
Available first on Google Cloud: Intel Optane DC Persistent Memory

A2 VMs now GA—the largest GPU cloud instances with NVIDIA A100 GPUs

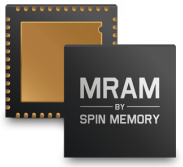
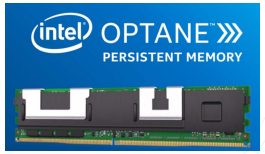
Personal Devices

Georgia Tech
CREATING THE NEXT

Heterogeneity Trade-offs



Source: memverge.com

Heterogeneity Characteristic	Emerging Technology	Hardware Vendors
Low Latency	MRAM	  Everspin Announces 1Gb ST-MRAM
High Bandwidth	HBM	 
Persistence	PMEM	

Georgia Tech
CREATING THE NEXT

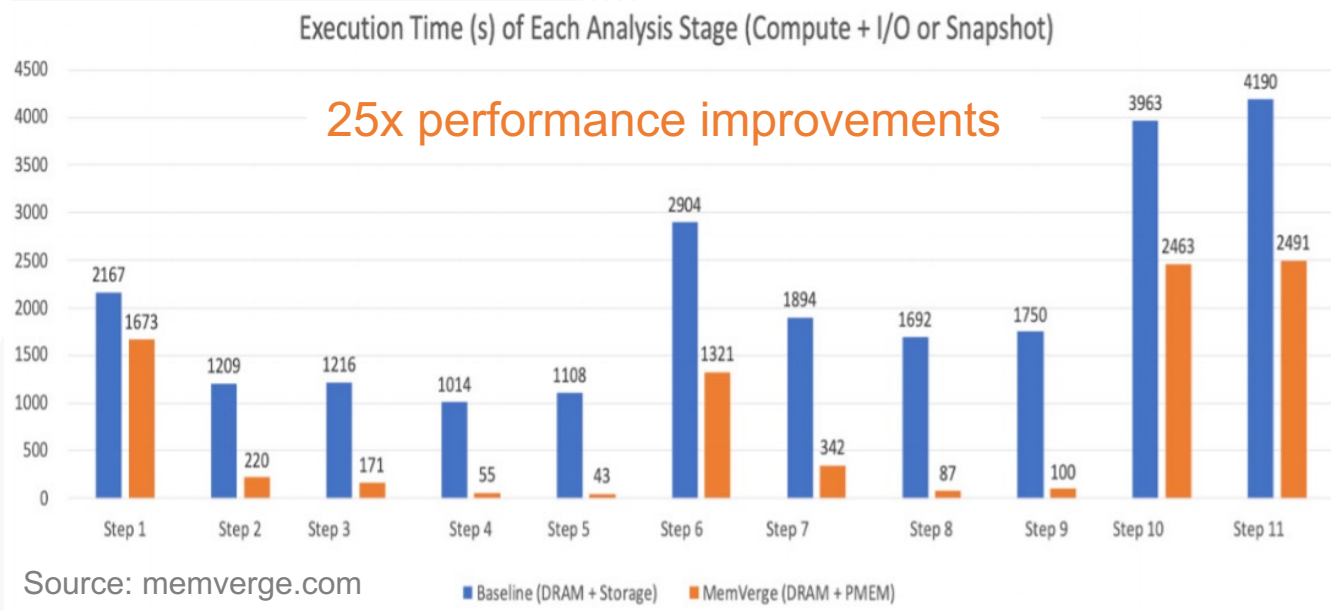
Real Impact on Real Applications

When using heterogeneous (hybrid) memories.

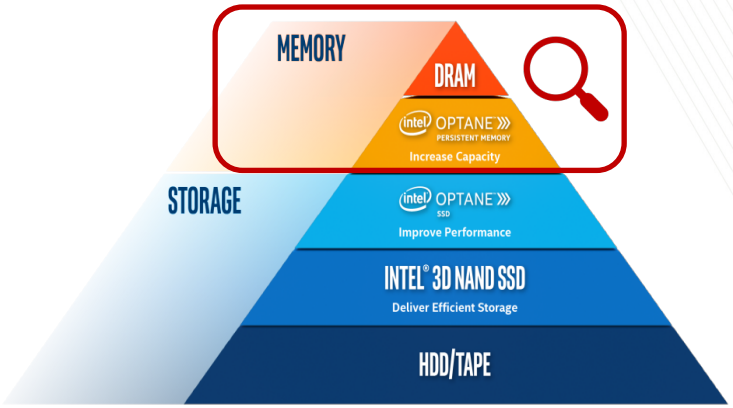
SOLUTION BRIEF

Big Memory Accelerates Single-Cell RNA Sequencing

MemVerge



Source: memverge.com



How to boost performance?

Dynamic Data Allocations
across the hardware layers.

Complex decision mix:

- Which / How much / Where / When to move data?
- Capacity sizing / sharing?

Georgia Tech
CREATING THE NEXT

Solutions across the Software Stack

Selective Publications.

Custom APIs for data allocation — Memif [ASPLOS '16], X-Mem [EuroSys '16] — Application Level

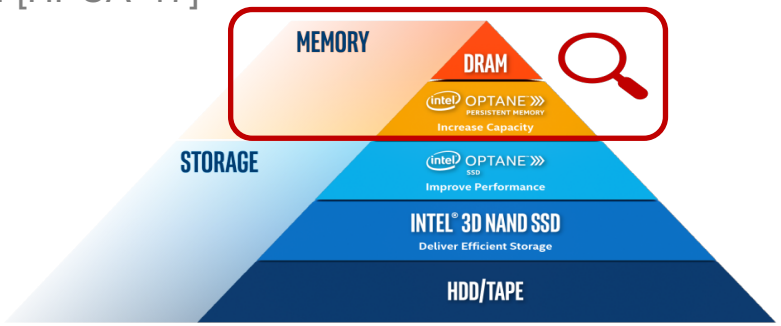
Runtime Directives — Unimem [SC '17], Tahoe [SC '18] — Runtime Level

Executable Instrumentation — Selective Page Migration [PACT '14] — Compiler Level

Page Scheduling — Thermostat [ASPLOS '17], Hetero-OS [ISCA '17], Nimble [ASPLOS '19], Leap [ATC '20] — Focus of my work — System Level

Hardware Level

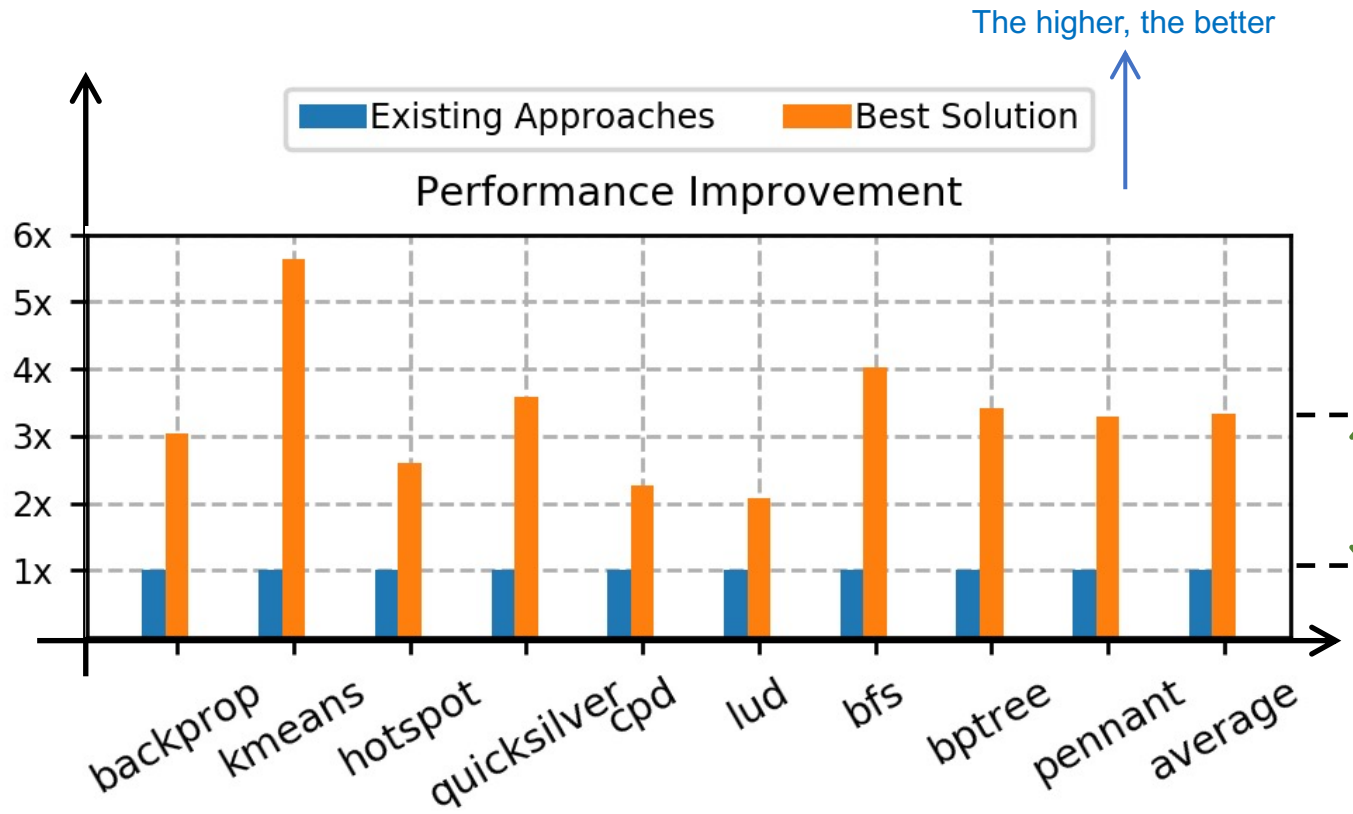
Specialized Hardware — MemPod [HPCA '17]



Georgia Tech
CREATING THE NEXT

Room for Performance Improvement

Left by existing approaches.



- A-priori knowledge
  - Data Access Patterns
- Fine-tuned operation
  - Extensive experimentation

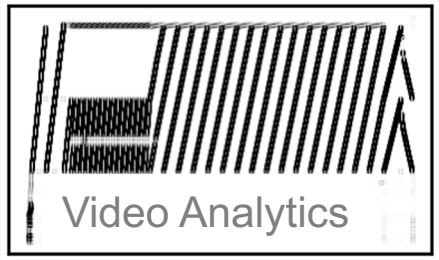
Best Solution

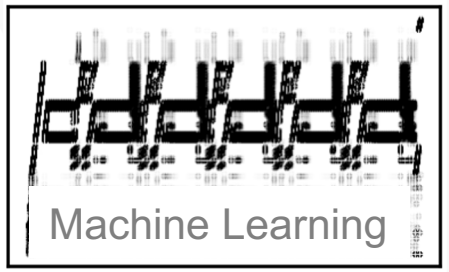
> 3x attainable improvement

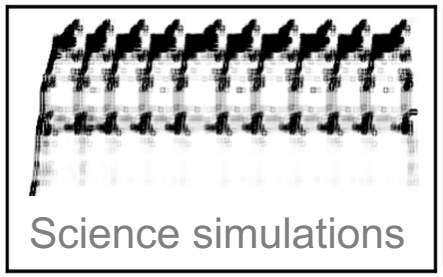
Existing Approaches

- Heuristics
  - Technology-specific
  - Application-specific
- Fixed configuration knobs
  - Empirically tuned

Research Contributions







Data access patterns

Applications

Resource Sharing — **CoMerge** - MEMSYS '17

$$$ *Cost Efficiency* — **Mnemo** - HPBDC '19

Operational Frequency Tuning — **Cori** – MEMSYS '20, IPDPS '21

Practical Machine Learning (ML) Integration

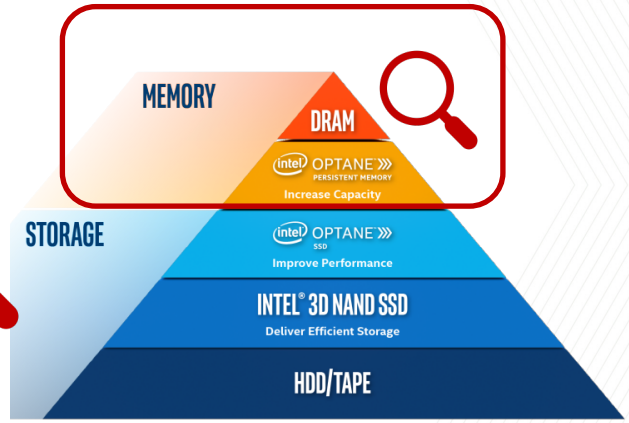
Design Foundations — **Kleio** – HPDC '19

Reducing ML Overheads — Under Submission

…to be continued

System-level Resource Manager

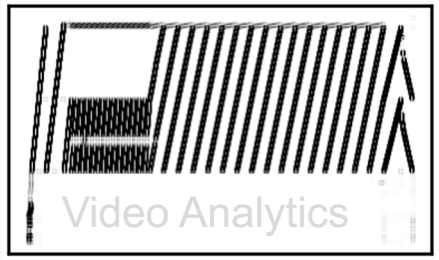
Hybrid Memory

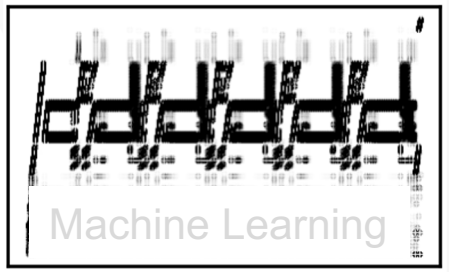


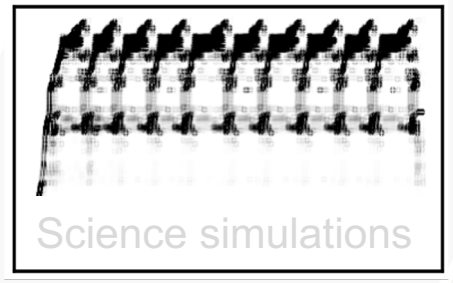
Heterogeneous Hardware

Georgia Tech
CREATING THE NEXT

Research Highlight







Data access patterns

Applications

Resource Sharing — **CoMerge** - MEMSYS '17

Cost Efficiency — **Mnemo** - HPBDC '19

Operational Frequency Tuning — **Cori** – MEMSYS '20, IPDPS '21

Practical Machine Learning (ML) Integration

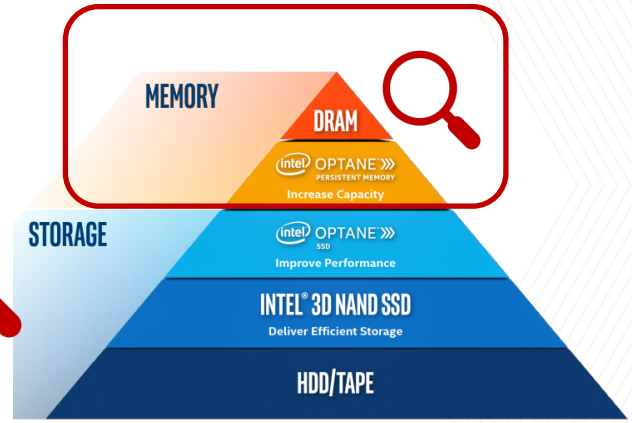
Design Foundations — **Kleio** – HPDC '19

Reducing ML Overheads — Under Submission

…to be continued

System-level Resource Manager

Hybrid Memory



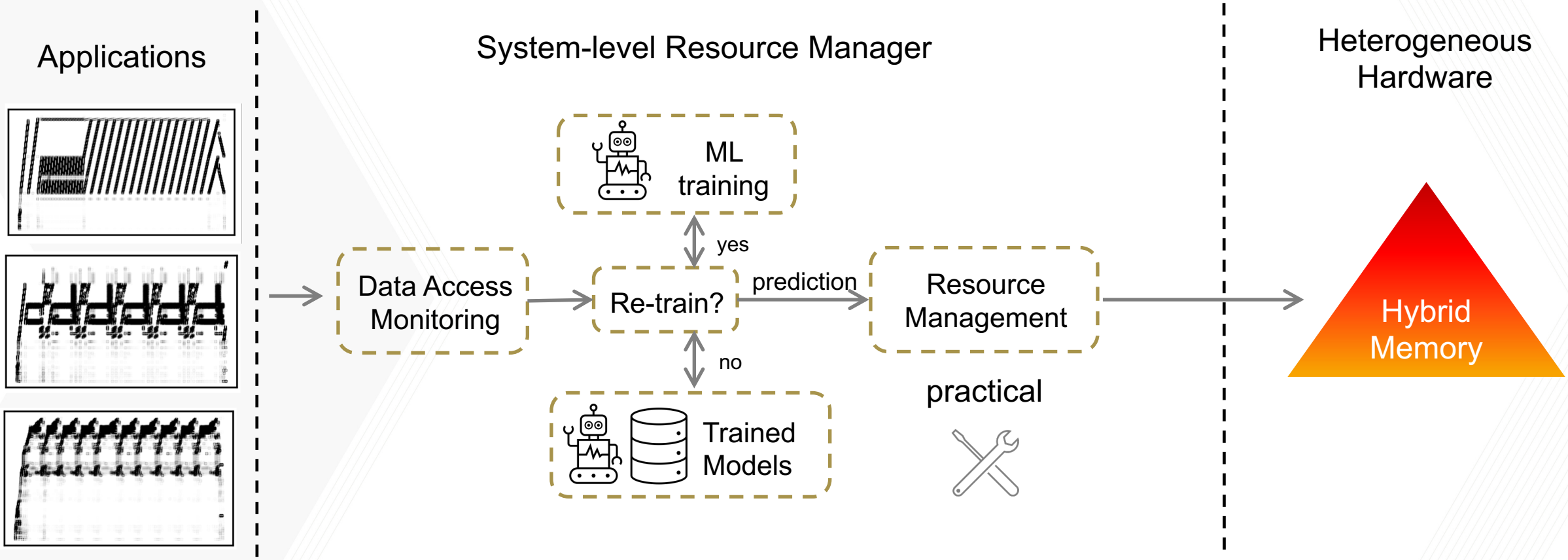
Heterogeneous Hardware

Georgia Tech
CREATING THE NEXT

Research Contributions
(Kleio)

The Goal

ML-augmented heterogeneous resource manager.

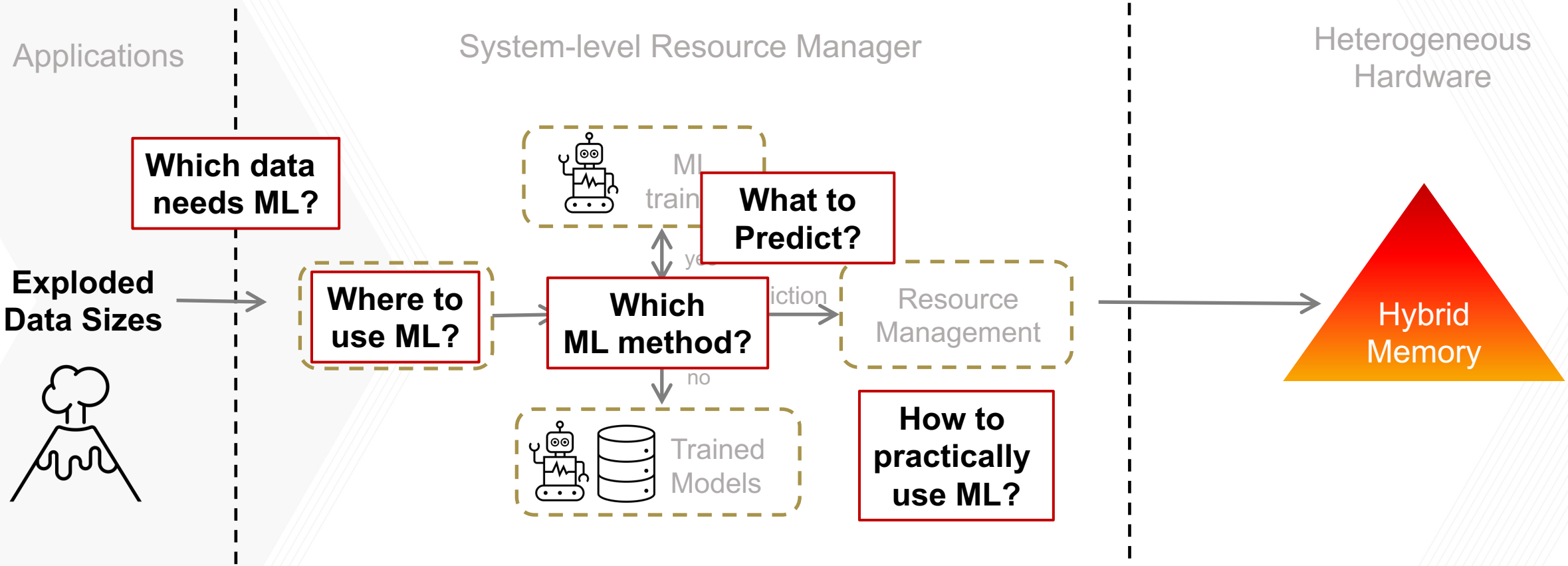


Who cares for such a solution?

- Hardware Vendors (e.g., AMD).
- Datacenters, Supercomputers.
- System for ML.

Contributions Towards the Goal

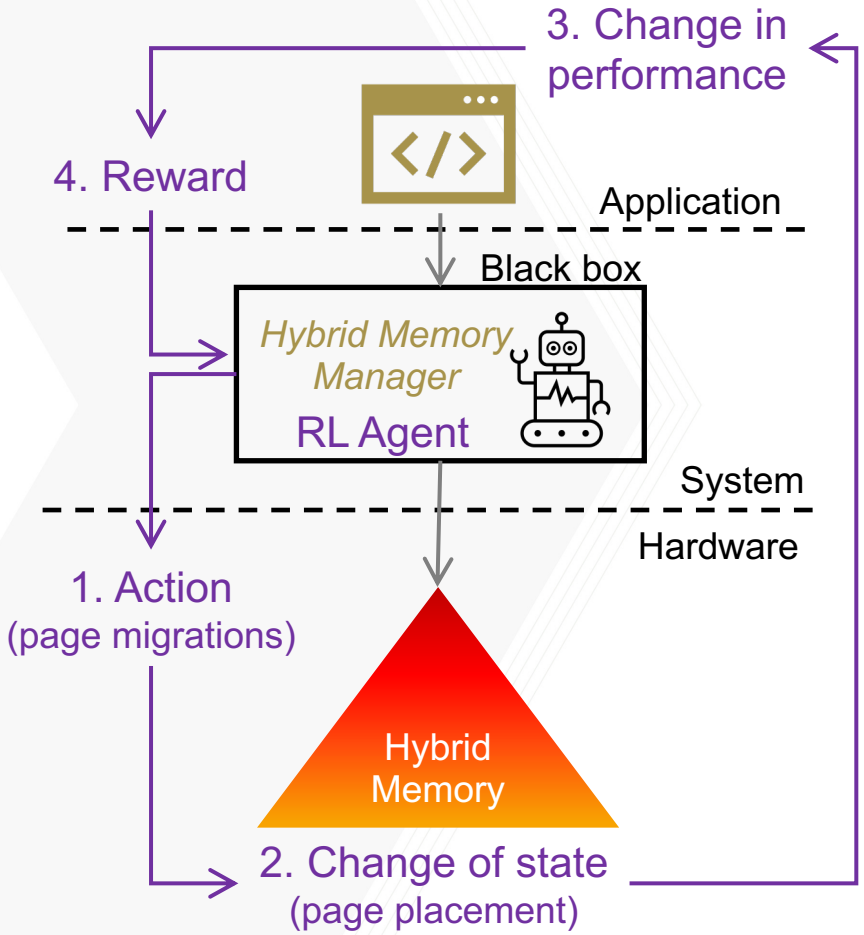
Laying the grounds for the *practical* integration of ML.



Georgia Tech
CREATING THE NEXT

Where to use ML?

Learn which pages to move. *Replace* the memory manager with ML.



Learn the Action: Learn from moving pages across hybrid memory using Reinforcement Learning. Learn from mistakes (e.g., cold pages in DRAM).

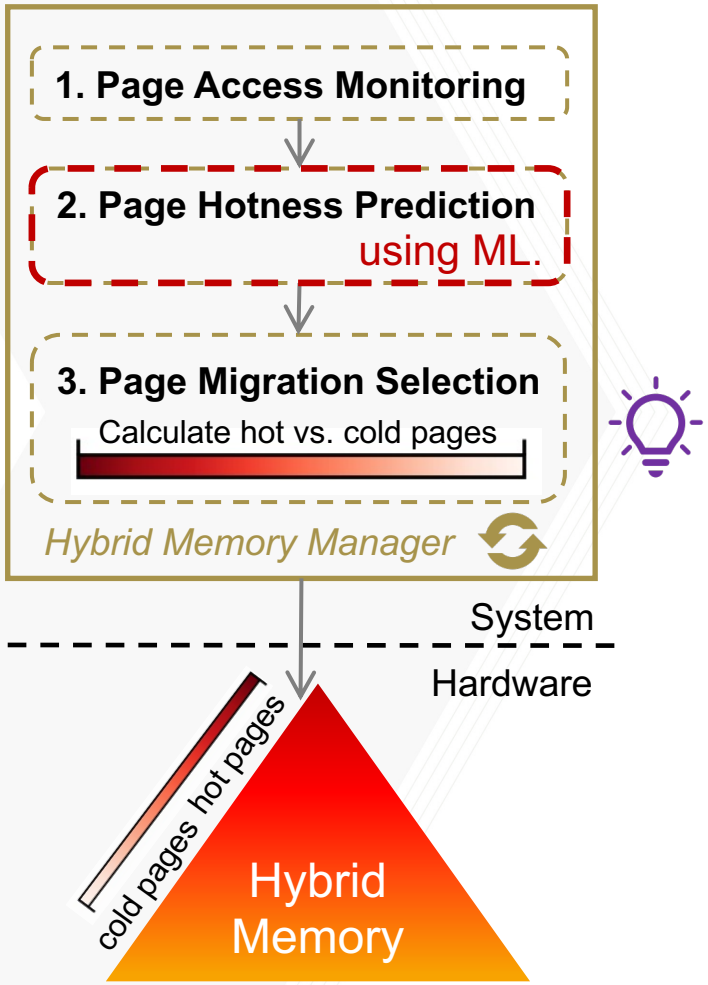
Why it is not a good fit:

- Exponential Action Space = 2^N
- Need to re-train if configuration of hybrid memory changes.
  - Number of memory units.
  - Difference in access speeds / capacities.

Not practical / scalable.

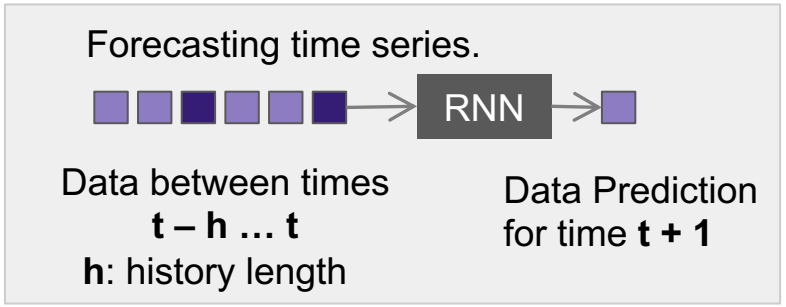
Where to use ML?

Learn which pages will be accessed in the future. *Augment* the memory manager with ML.



Learn the Behavior: Learn which pages will be accessed in the future. The manager will then move hot and cold pages appropriately.

Recurrent Neural Networks

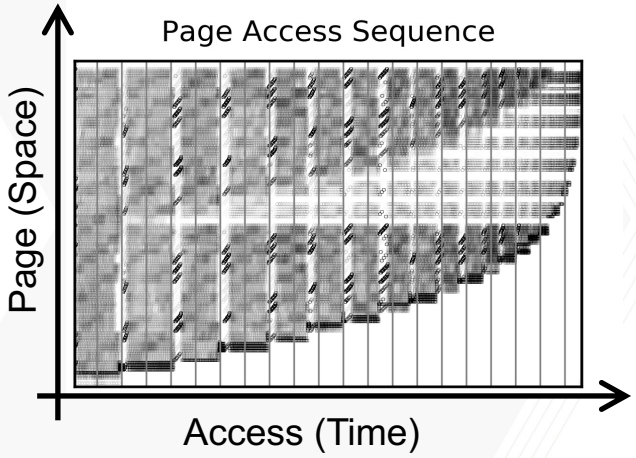


What to Predict with RNNs?

Next page accessed vs. page hotness.

Exploded Data Sizes

Memory access trace = Time series of memory access.



Learn which page will be accessed next.

Page Access Sequence
e.g., 100, 101, 102..

→ RNN →

Next Page to be Accessed
e.g., 103

No. models	Overheads	Accuracy
1 per app	Days to train. Months to fine-tune.	Low. Top-k predictions not useful.

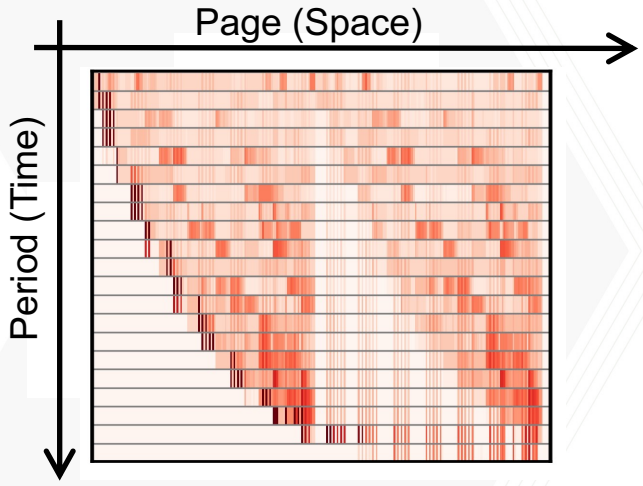
Georgia Tech
CREATING THE NEXT

What to Predict with RNNs?

Next page accessed vs. page hotness.

Exploded Data Sizes

Flip the view of the problem!



Page Access Hotness per Period

Learn how hot a page will be in the future.

Page Access Hotness Across previous Periods e.g., 100, 0, 0, 100.. → RNN → Page Access Hotness in the next Period. e.g., 0

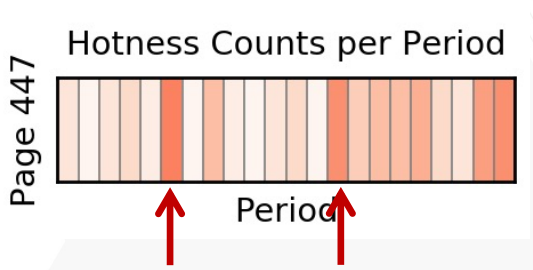
No. models	Overheads	Accuracy
Many per app	Parallel training. Smaller models.	High accuracy.

How to practically use ML?
Augment existing approaches with ML.

Greek Trivia: According to the ancient Greek mythology, Kleio was the muse of history, daughter of Mnemosyne, goddess of memory.

[HPDC '19] ***Kleio:*** *a Hybrid Memory Page Scheduler with Machine Intelligence.*

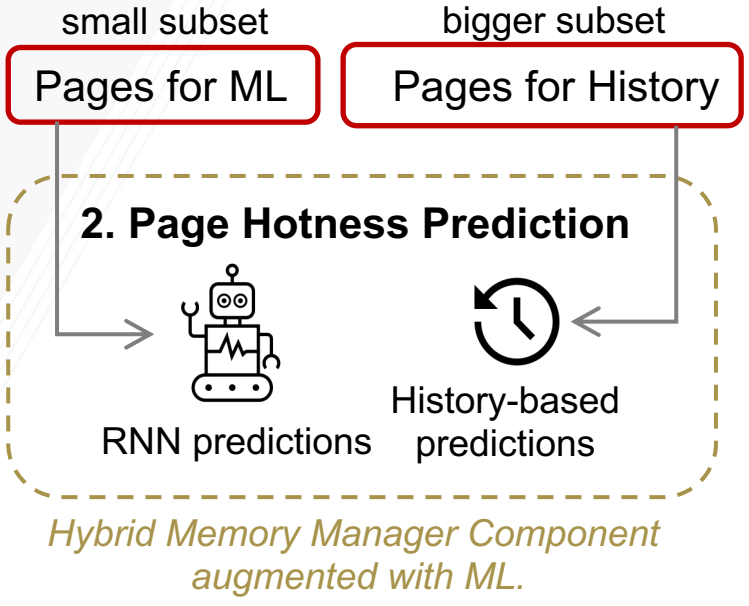
Do all pages need ML? No! Only the ones that current history-based solutions manage inefficiently.



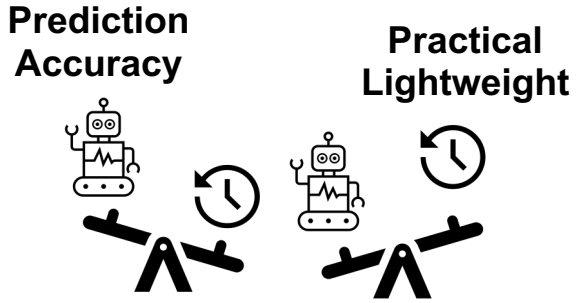
Suddenly becomes very **hot**.
Misplaced (not in DRAM when hot).

Small subset of such pages.

Key Idea



Hybrid Memory Manager Component augmented with ML.



Evaluation

Kleio delivers on average 80% of the attainable performance improvements.

100% Accurate ML
for Selected Pages

History-based
Rest of Pages

"Best" Solution

RNNs for
Selected Pages

History-based
Rest of Pages

Kleio

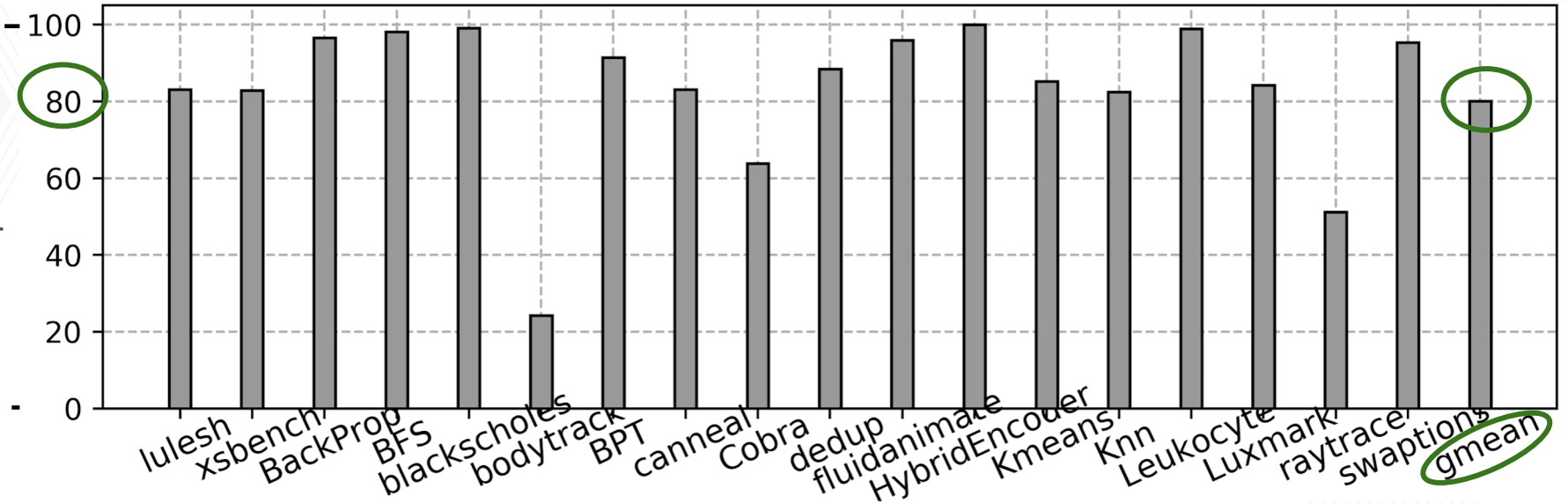
Baseline Solution

History-based
Predictions for all Pages

Attainable performance

The higher
The better

More than **95%** for **half** of the applications!

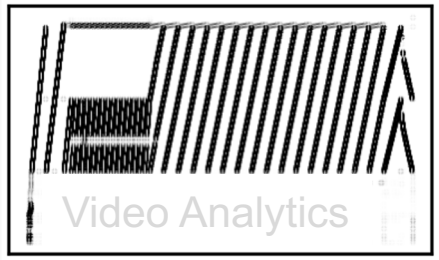


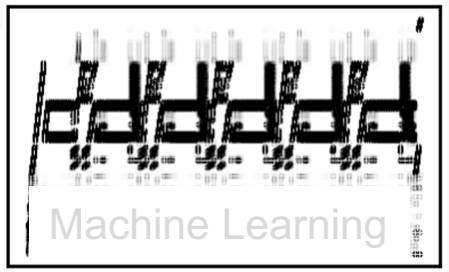
For 100 selected pages.

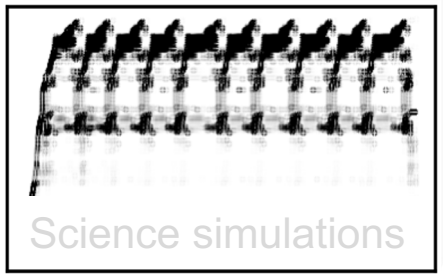
Georgia Tech
CREATING THE NEXT

Research Contributions
(Other)

Can we do better?







Data access patterns

Applications

Resource Sharing **CoMerge** - MEMSYS '17

$$$ *Cost Efficiency* **Mnemo** - HPBDC '19

Operational Frequency Tuning **Cori** – MEMSYS '20, IPDPS '21

Practical Machine Learning (ML) Integration

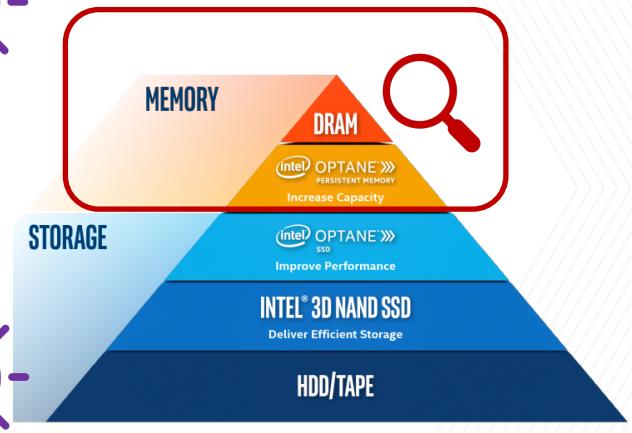
Design Foundations **Kleio** – HPDC '19

Reducing ML Overheads Under Submission

…to be continued

System-level Resource Manager

Hybrid Memory



Heterogeneous Hardware

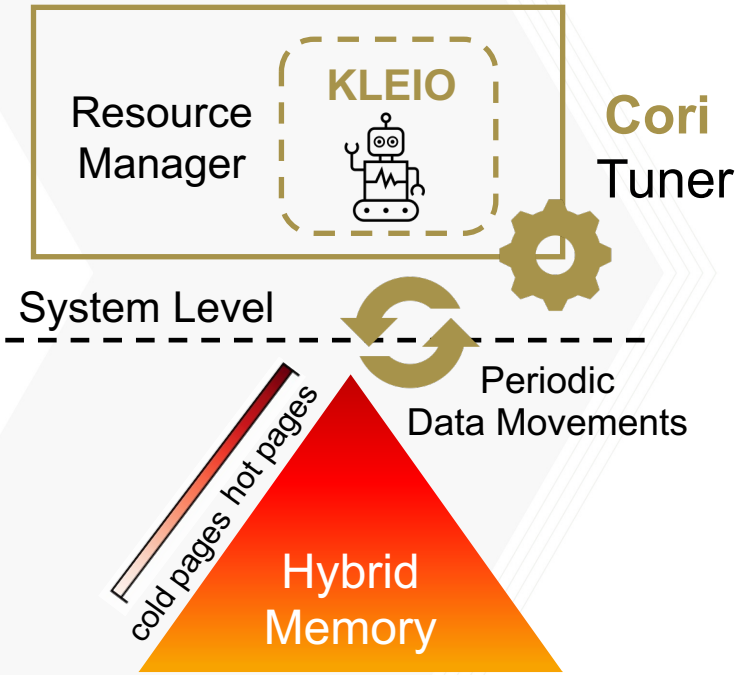
Georgia Tech
CREATING THE NEXT

Boosting the Effects of Machine Learning

Using observation-driven insights.

Greek Trivia: According to the ancient Greek mythology, Cori (short for Terpsichore) was the muse of dance, sister of Kleio, daughter of Mnemosyne, goddess of memory.

[IPDPS '21] ***Cori:*** *Dancing to the Right Beat of Periodic Data Movements over Hybrid Memory Systems.*



Cori tunes the frequency of data movements.

Data Placement Benefit

Data Movement Cost

Balance

Performance boost.

Key Idea

Data Reuse Times

Data Movement Periods

Synchronize

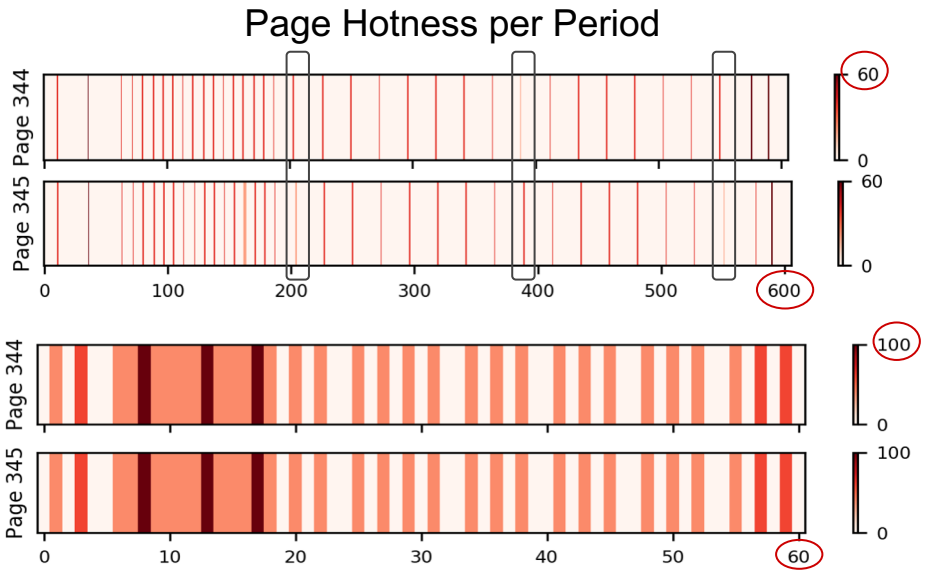
Georgia Tech
CREATING THE NEXT

It Is All About The Right Granularity

[Under Submission] *Clustering Patterns for Practical Machine Intelligent Hybrid Memory Management.*

Tuning the resource manager's operational frequency (period) …

… tunes the patterns for ML!



Kleio
Sequences are slightly different.

Cori
Sequences are **identical**.

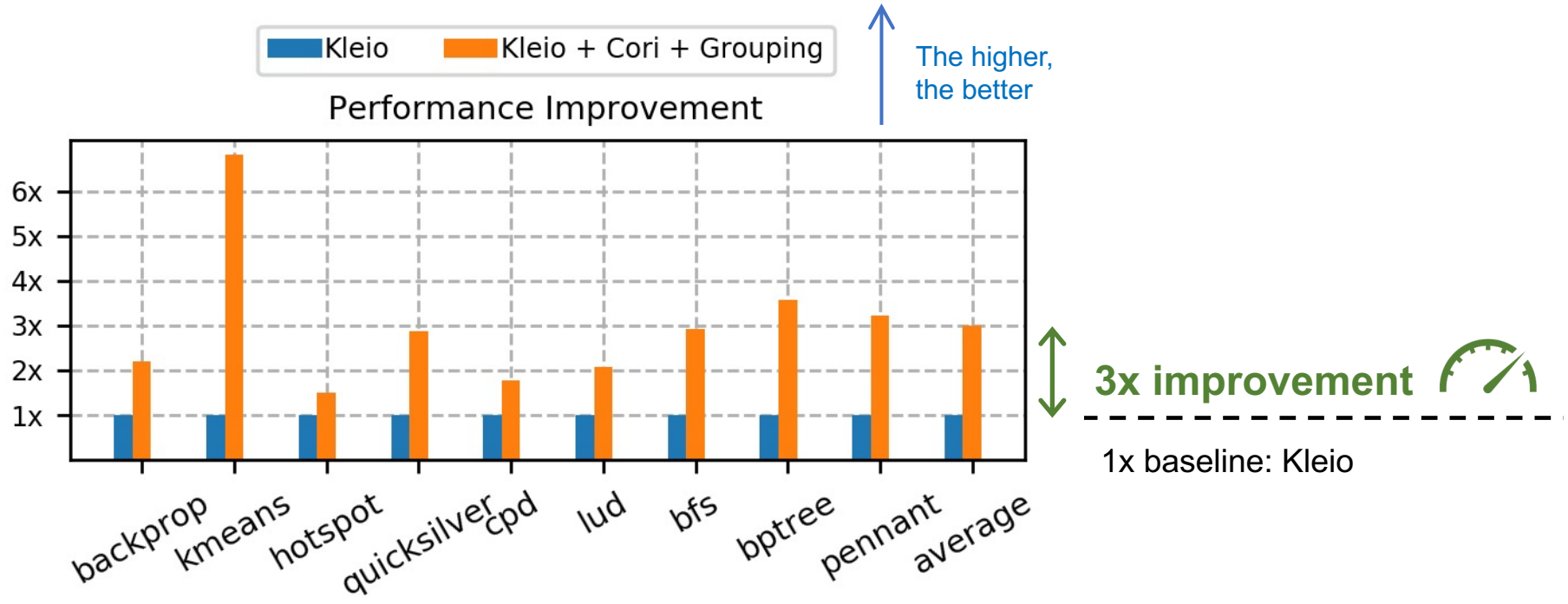
Key Idea

Group pages with *identical* patterns under a single ML model.

↓ ML overheads

Georgia Tech
CREATING THE NEXT

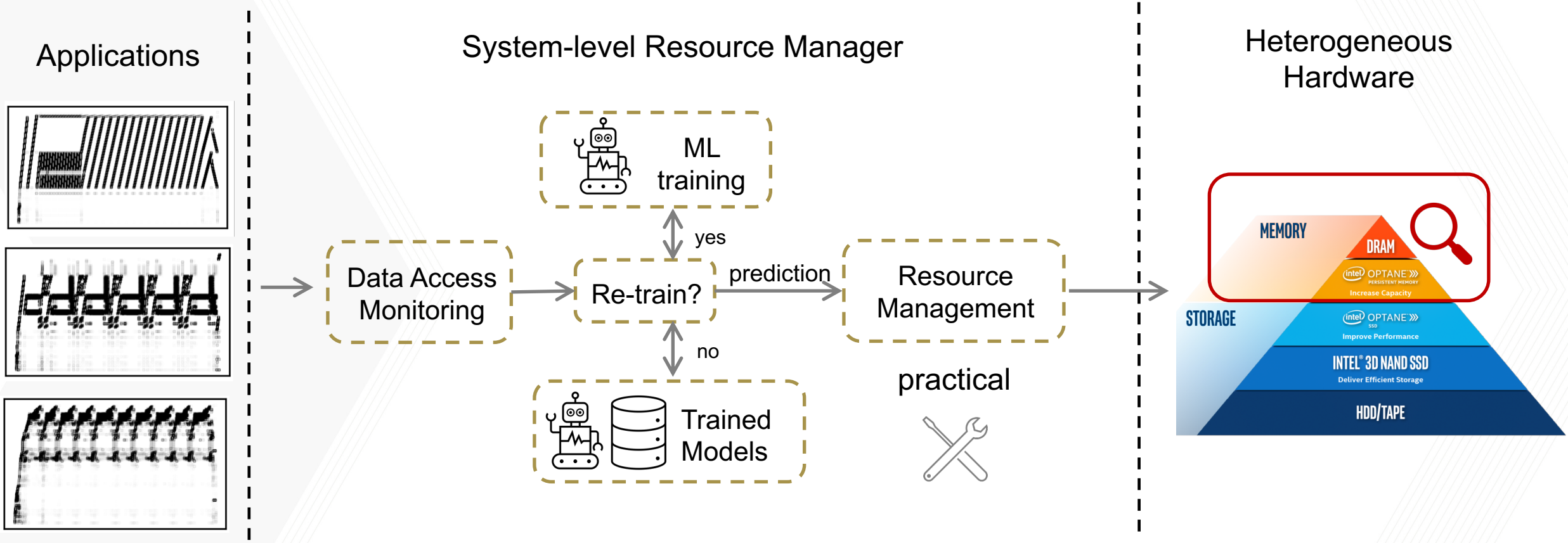
Boosting Application Performance



Fine-tuned operation further boosts the effects of ML in resource management.

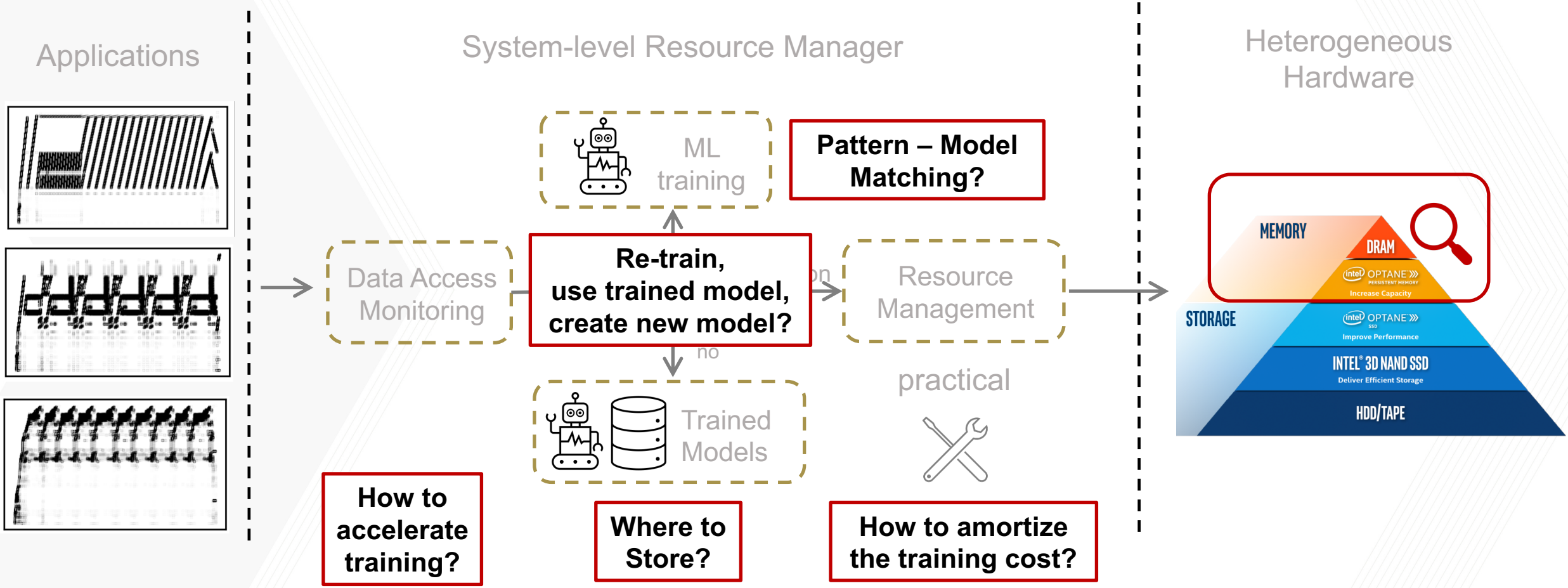
Future Research Directions

ML-augmented Heterogeneous Resource Manager



Immediate Future Contributions

Fully integrated adaptive resource manager.



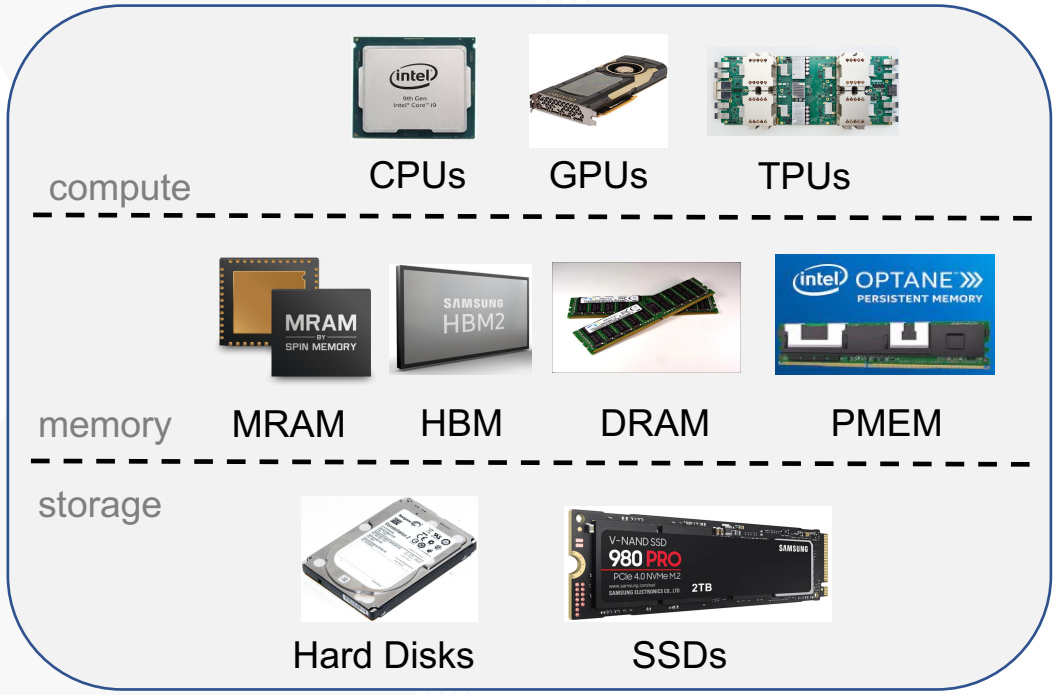
Georgia Tech
CREATING THE NEXT

Intelligent Management of Extreme Heterogeneity

Hardware configuration?
Data / Resource Management across layers / nodes?

Users

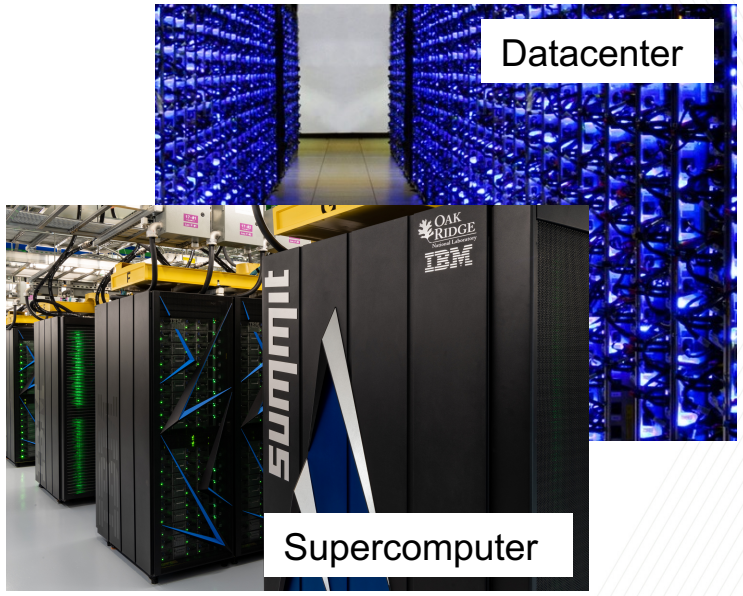
Multi-tenancy?
Isolation?



Local Node

Performance?
Cost / Energy / Resource Efficiency?

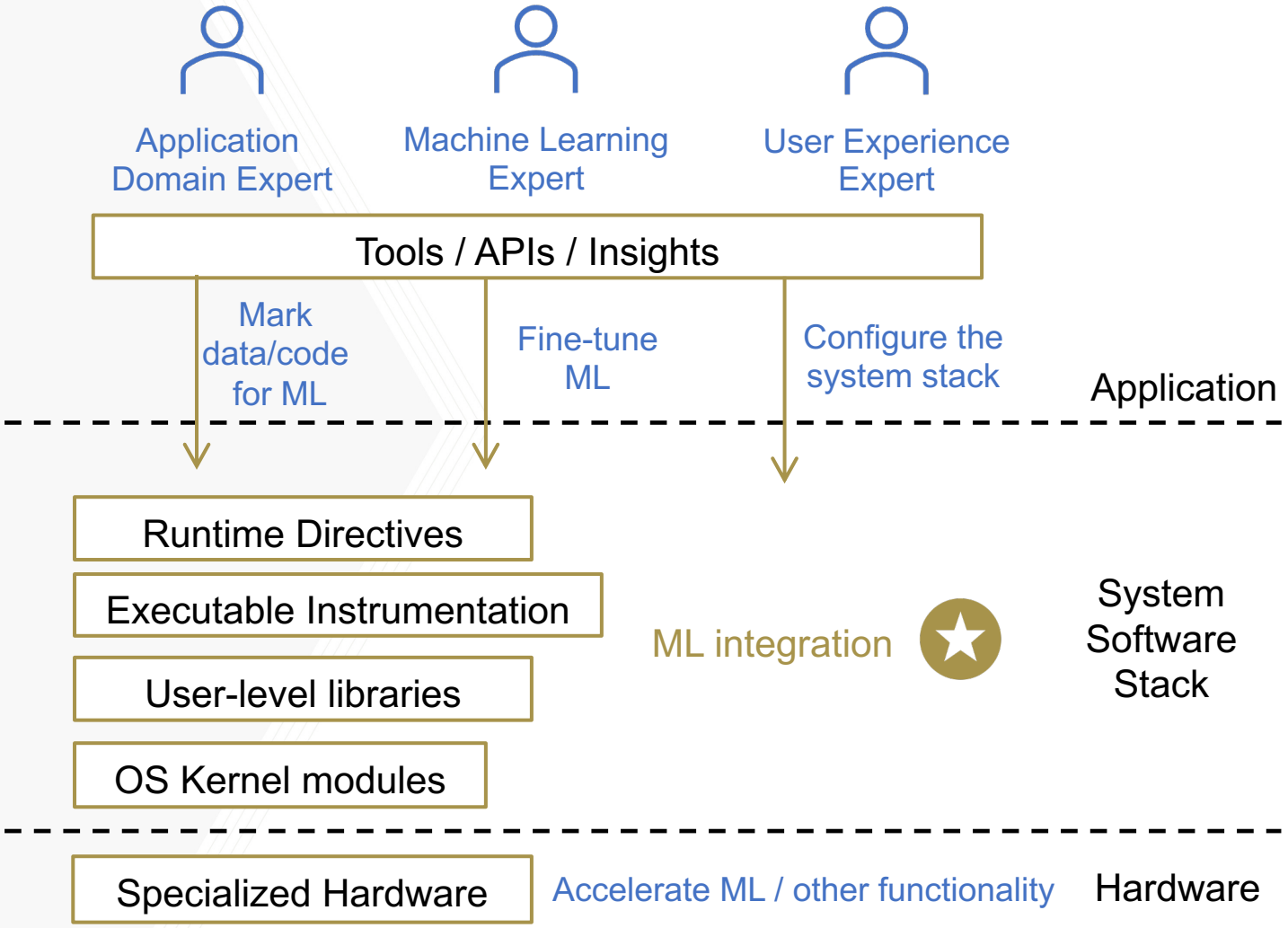
High-Speed Interconnects



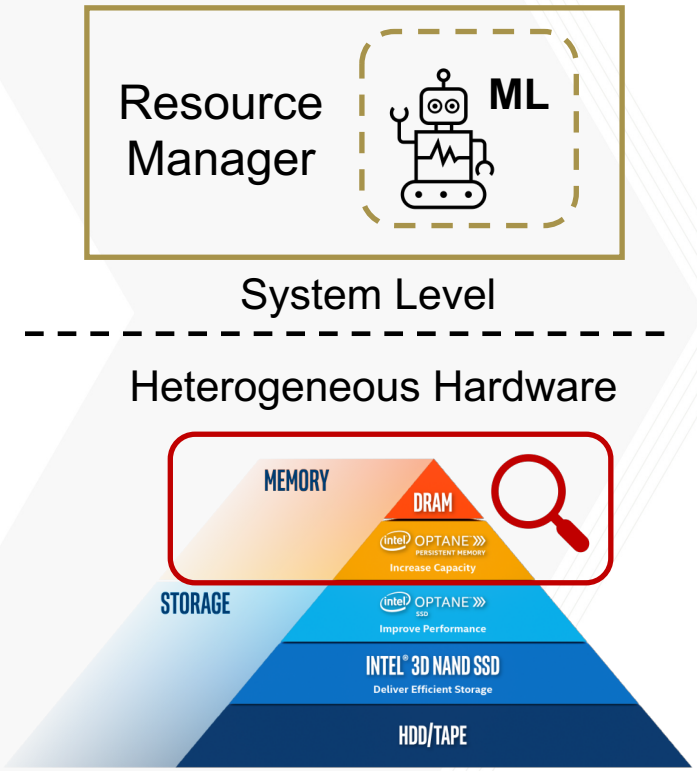
Massive Node Clusters
Disaggregated Resources

ML integration Aspects:
Necessity Effectiveness Practicality Interpretability

Cross-Stack Synergies for ML integration



Summary



Greek Trivia: According to the ancient Greek mythology, **Thaleia** was the muse of comedy, daughter of Mnemosyne, goddess of memory.

First-author Publications

Resource Sharing — **CoMerge** - MEMSYS '17

Cost Efficiency — **Mnemo** - HPBDC '19

Operational Frequency Tuning — **Cori** – MEMSYS '20, IPDPS '21

Practical Machine Learning (ML) Integration

Design Foundations — **Kleio** – HPDC '19

Reducing ML Overheads — Under Submission

…to be continued

Mnemosyne

Georgia Tech
CREATING THE NEXT