Building Smart and Fast Systems using Machine Learning and Computer Vision.

Thaleia Dimitra Doudali
Assistant Research Professor @IMDEA Software Institute
Born and raised in Greece.

Undergrad in ECE at NTUA, Athens, Greece.

PhD in CS at Georgia Tech, Atlanta, USA.

Assistant Professor at IMDEA, Madrid, Spain.

Advised by Ada Gavrilovska.
My research lies at the intersection of Machine Learning and Systems.

**ML for Systems**
- e.g., RNNs for system-level pattern prediction.

**Systems for ML**
- e.g., design new systems to optimize ML workloads.

**ML + CV for Systems**
- e.g., image-based ML for pattern recognition and prediction.
Why do we need Smarter and Faster Systems?
The evolution of the hardware technologies, calls for software improvements.

Building Smart Systems
Using machine and human intelligence to build practical ML-based systems.

Building Fast Systems
Reducing ML-based management overheads with visualization.
Building image-based system pipelines.

Future Research Directions
Talk Outline

Why do we need Smarter and Faster Systems?
The evolution of the hardware technologies, calls for software improvements.

Building Smart Systems
Using machine and human intelligence to build practical ML-based systems.

Building Fast Systems
Reducing ML-based management overheads with visualization.
Building image-based system pipelines.

Future Research Directions
The Era of Data

“More than 65 ZB of data will be created, captured, copied, and consumed in the world this year.”

Source: International Data Corporation, March 2021.

**Data Analytics Pipeline**

ZBs of data

Capture | Process | Store | Analyze | Use

Need for speed and massive storage capacities!
The Era of Heterogeneous Hardware

**Compute Acceleration**
- Nvidia GPUs
- Cloud TPU v2
  - 180 teraflops
  - 64 GB High Bandwidth Memory (HBM)

**Data Storage Acceleration**
- Intel OPTANE Persistent Memory
- High Bandwidth Memory

**Network Acceleration**
- Mellanox Innovu™-2 Flex
  - Open Programmable SmartNIC

**Interconnection Standards**
- CXL (Compute Express Link™)
  - Gen-Z Consortium
  - Industry Leaders developing a memory-semantic interconnect
Heterogeneity Across Computing Platforms

**Supercomputers**

**Exascale Era**

**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**

Sweden's National Supercomputer Center Launches 300 AI Dataflow System - "Berzellius"

By Staff A

<table>
<thead>
<tr>
<th>Application Performance</th>
<th>200 PF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Nodes</td>
<td>4,608</td>
</tr>
<tr>
<td>Node performance</td>
<td>42 TF</td>
</tr>
<tr>
<td>Memory per Node</td>
<td>512 GB DDR4 + 96 GB HBM2</td>
</tr>
<tr>
<td>NV memory per Node</td>
<td>1600 GB</td>
</tr>
<tr>
<td>Total System Memory</td>
<td>&gt;10 PB DDR4 + HBM2 + Non-volatile</td>
</tr>
<tr>
<td>Processors</td>
<td>2 IBM POWER9™ 9,216 CPUs 6 NVIDIA Volta™ 27,648 GPUs</td>
</tr>
<tr>
<td>File System</td>
<td>250 PB, 2.5 TB/s, GPFS™</td>
</tr>
<tr>
<td>Power Consumption</td>
<td>13 MW</td>
</tr>
<tr>
<td>Interconnect</td>
<td>Mellanox EDR 100G InfiniBand</td>
</tr>
<tr>
<td>Operating System</td>
<td>Red Hat Enterprise Linux (RHEL) version 7.4</td>
</tr>
</tbody>
</table>
Heterogeneity Trade-offs

We are in the era of **Hybrid Memory** Systems.
A mix of different technologies at different speeds / capacities / costs.
The OS should move pages dynamically across hybrid memory to maximize the efficiency.
Need for Smarter Hybrid Memory Management

It is a complex decision mix to manage the data allocated across memories.

E.g., Which / How much / Where / When to move data?

Why do we need smarter and faster systems?

- Application data sizes
- Complex data access patterns
- Exploded system parameter space

Hard to balance

practicality

performance
Building Smart Systems
Using machine and human intelligence to build practical ML-based systems.

Building Fast Systems
Reducing ML-based management overheads with visualization.
Building image-based system pipelines.

Future Research Directions

Why do we need Smarter and Faster Systems?
The evolution of the hardware technologies, calls for software improvements.
The Vision
ML-augmented heterogeneous resource manager.
Contributions Towards the Vision
Laying the grounds for the *practical* integration of ML.
System design of Kleio

Kleio: a hybrid memory page scheduler with machine intelligence. [HPDC 2019]

1. Page Access Monitoring

2. Page Hotness Prediction
   - ML-based predictions (Per page RNN models)
   - History-based predictions

3. Page Migration Selection
   - Calculate hot vs. cold pages

Result: Kleio bridges 80% of the performance gap between existing and oracular solutions.

Not all pages “need” ML.
Apply ML **when** and **where** necessary.

Apply ML on a small page subset.
- **Foundations for practical use of ML.**

Carefully select pages for ML.
- **Application performance boost.**

Can we reduce the number of pages via clustering?

Small can still mean thousands of pages, because of the massive memory footprints of modern workloads.
**Insights from the System Design of Coeus**

**Coeus:** Clustering (A)like Patterns for Practical Machine Intelligent Hybrid Memory Management . [CCGrid 2022]

---

**Clustering?** Let’s use ML!

For example, K-means.
- How many clusters?
- Clustered input to ML?

Not trivial to configure.

Let’s use our human intelligence..

.. Kleio learns the patterns of page hotness across time periods.

**Key Idea**

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

---

So what if I increase the duration of the period?

↓ 3x less RNNs

↑ 3x more performance
Why do we need Smarter and Faster Systems?
The evolution of the hardware technologies, calls for software improvements.

Building *Smart* Systems
Using machine and human intelligence to build practical ML-based systems.

Building *Fast* Systems
Reducing ML-based management overheads with visualization.
Building image-based system pipelines.

Future Research Directions
The Key(s) to a Practical and Efficient ML-based System Design

Apply ML **when** and **where** necessary.

- **Kleio’s Page Selector**
  - Priority
  - Page hotness predictions
  - History
  - Pool of Pages

Apply ML on a small page subset.

- Foundations for practical use of ML.

Carefully select pages for ML.

- Application performance boost.

The page selection is not a lightweight process. Performance modeling and estimations are used to maximize the effects of ML on application performance.

Can we accelerate the page selection process?
Insight from Visualizing Pages Selected for ML

Neighboring pages that are part of distinct access patterns across time receive similar priority for ML.
Towards Image-based Page Selection

**Cronus**: Computer Vision-based Machine Intelligent Hybrid Memory Management. [MEMSYS 2022]

- **1. Image Creation**
- **2. Pattern Detection**
- **3. Page Selection**

Cronus reduces by 400x the page selection times, from minutes down to seconds.
Creating images helps:

- Another way to represent data, reducing their dimensionality to a 2D / 3D space.
- Captures spatial and temporal correlations.
- Leverage computer vision and image-based algorithms.

**Feature Extraction**

**Image-based ML Classifiers**

**Earthquake Detection:**
Extract Frequencies of Seismic Waves.

**Stock Market Forecasting:**
Trading by learning time series data as images.

**Autonomous Driving:**
Object Detection & Recognition
What can an image-based system pipeline look like?

E.g., predicting future resource utilization.

- Input Image
- Pattern Recognition (e.g., memory utilization)
  - Choose based on pattern.
- Pre-trained ML models
  - Pattern Prediction
- Class = “sinusoid”
  - Prediction
- Not Accurate Prediction? Retrain.

Resource Management System
Image-based vs. Number-based Machine Learning

Research paper under submission.

Image-based ConvLSTM model

The image-based ConvLSTM makes more accurate predictions.
What can an image-based system pipeline look like?

E.g., learning memory access patterns.

Pattern Recognition

Choose based on pattern.

Pattern Prediction

Not Accurate Prediction? Retrain.

Pre-trained ML models

Memory Manager

System Level

Hardware Level

Hybrid Memory

image + metadata

Class = “stride”

E.g., learning memory access patterns.

Not Accurate Prediction? Retrain.
Early Results on Image-based Pattern Prediction

Ground Truth

Prediction

More challenging, since the data access patterns are more complex.
Why do we need Smarter and Faster Systems?
The evolution of the hardware technologies, calls for software improvements.

Building *Smart* Systems
Using machine and human intelligence to build practical ML-based systems.

Building *Fast* Systems
Reducing ML-based management overheads with visualization.
Building image-based system pipelines.

Future Research Directions
Future Research Directions

My research lies at the intersection of Machine Learning and Systems.

- **ML for Systems**: E.g., Online practical training, ML for different systems problems.
- **Systems for ML**: E.g., Optimize memory management for RNNs / ML workloads.
- **ML + CV for Systems**: E.g., Image-based pattern recognition and prediction of resource usage.

---

**Operating Systems (OS)**

**Software**
Intelligent Management of Extreme Heterogeneity

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

Users

Multi-tenancy?  
Isolation?

Performance?  
Cost / Energy /  
Resource Efficiency?

High-Speed Interconnects

Datacenter

Supercomputer

Massive Node Clusters  
Disaggregated Resources

Local Node

compute
CPUs  GPUs  TPUs

memory
MRAM  HBM  DRAM  PMEM

storage
Hard Disks  SSDs

System vs. HW / SW co-design?

ML integration Aspects:  
Necessity  Effectiveness  Practicality  Interpretability
How can we use our human intelligence to build **practical** systems that leverage machine learning and computer vision?