Towards ML-driven resource orchestration in disaggregated memory systems: Challenges and Opportunities

Dimosthenis Masouros

National Technical University of Athens, Greece

May 19th 2023

Invited Talk – 2nd Workshop on Composable Systems

Co-located with IPDPS 2023

"ToC"

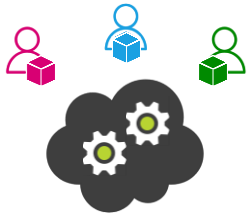
ML for (disaggregated) systems

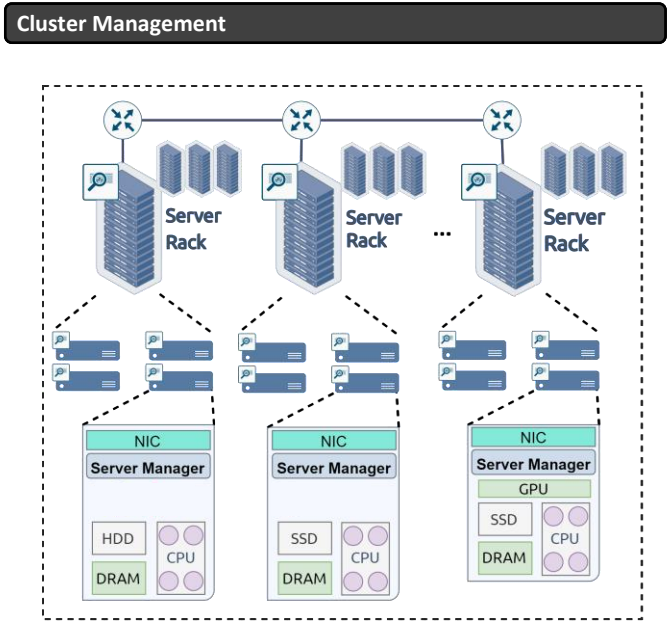
Adrias

Interference-Aware Memory Orchestration for Disaggregated Cloud Infrastructures

Introduction

- Rise of applications executed in the Cloud
- Application co-location (multi-tenancy)
  - Resource interference ➔ performance degradation
- Traditional infrastructures ➔ static architecture
  - Servers with fixed number of CPUs and RAM + HW accelerators
- Several issues/challenges w.r.t. resource efficiency
  - Fragmentation of resources
  - Handling of HW failures
  - Integration of new HW devices



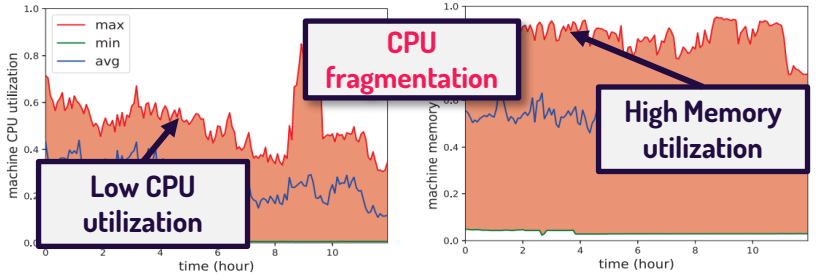


Why Memory Disaggregation?

- Memory is a major factor w.r.t. resource fragmentation
  - Memory shortage
  - Memory stranding
- Memory disaggregation to the rescue!
  - Allocate memory either locally or from neighboring nodes over network

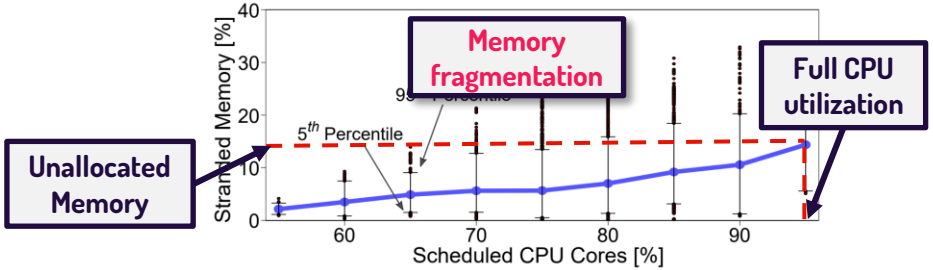
Imbalance in the cloud: An analysis on Alibaba cluster trace

Chengzhi Lu, Kejiang Ye, +2 authors Tongxin Bai • Published 1 December 2017 • Computer Science •
2017 IEEE International Conference on Big Data (Big Data)



Pond: CXL-Based Memory Pooling Systems for Cloud Platforms

Huaicheng Li, Daniel S. Berger, +10 authors R. Bianchini • Published 1 March 2022 • Computer Science •
Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems



Why Memory Disaggregation?

- Memory is a major factor w.r.t. resource fragmentation
  - Memory shortage
  - Memory stranding

- Memory disaggregation to the rescue!
  - Allocate memory either locally or from neighboring nodes over network
  - Memory disaggregation and multi-tier memory architectures infiltrating the Cloud world

Towards an Adaptable Systems Architecture for Memory Tiering at Warehouse-Scale

Padmapriya Duraisamy, Wei Xu, +15 authors A. Vahdat • Published 25 March 2023 • Computer Science •
Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems

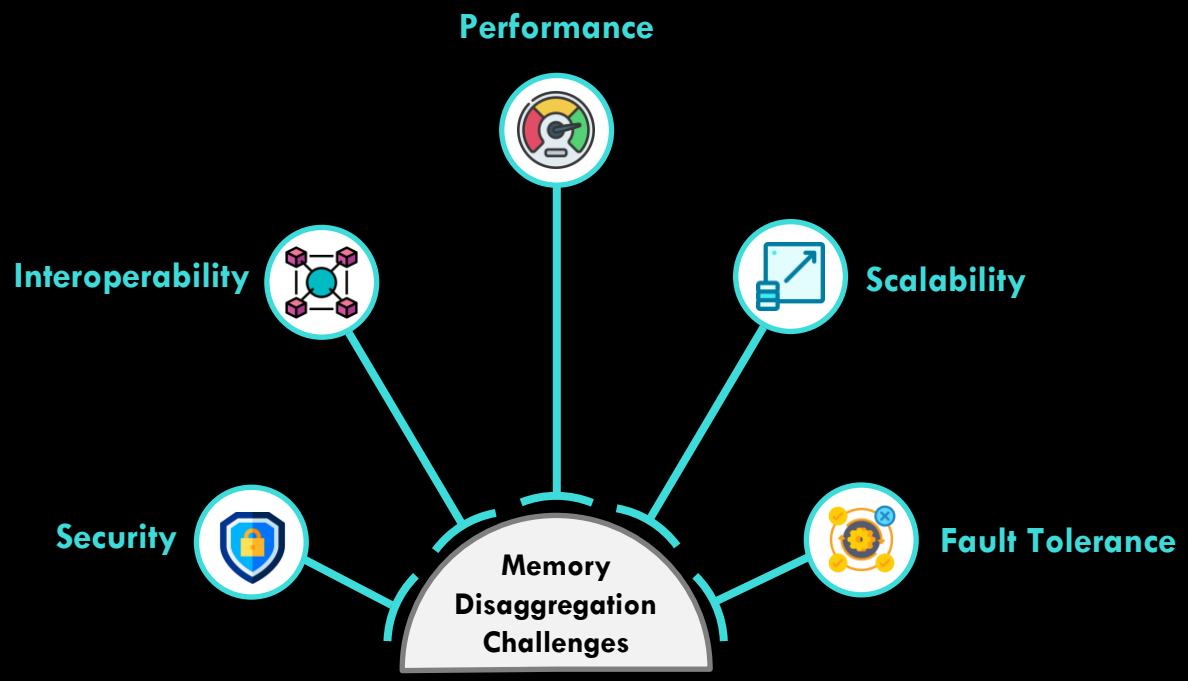
Design Tradeoffs in CXL-Based Memory Pools for Public Cloud Platforms

Daniel S. B

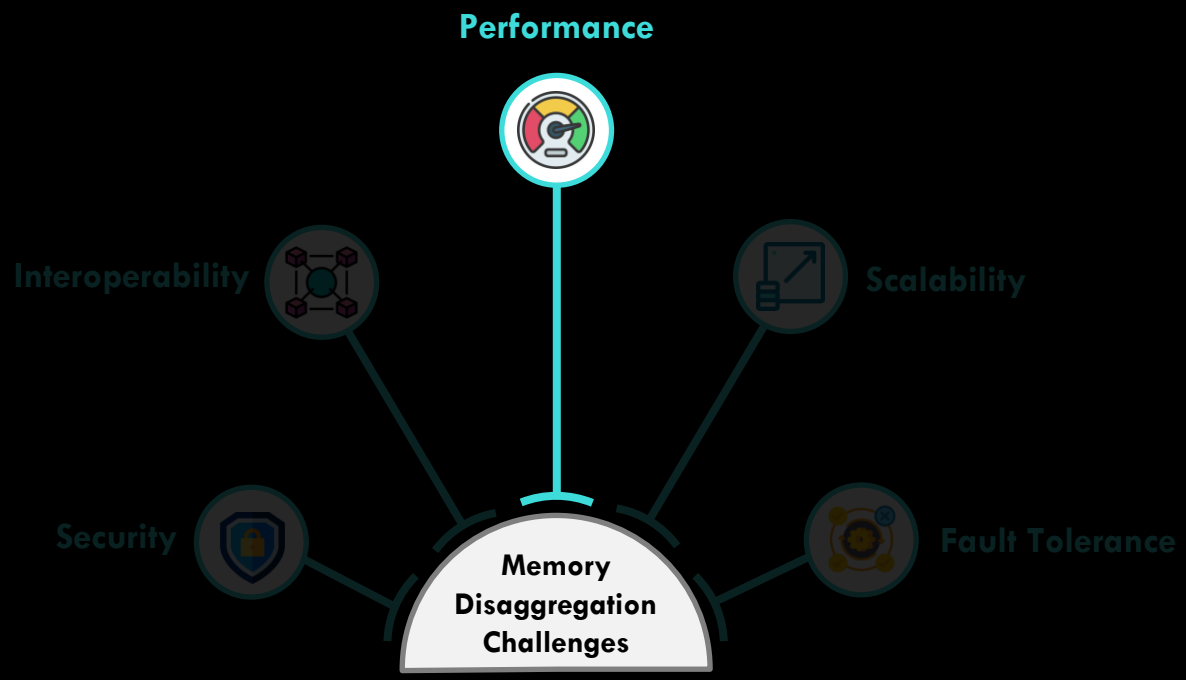
Pond: CXL-Based Memory Pooling Systems for Cloud Platforms

Huaicheng Li, Daniel S. Berger, +10 authors R. Bianchini • Published 1 March 2022 • Computer Science •
Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems

Challenges of Memory Disaggregated Cloud systems



Challenges of Memory Disaggregated Cloud systems



Performance Implications of Memory Disaggregation

- Performance degradation by design **(OR NOT?)** 🤔
  - Network Latency/Bandwidth & Protocol overheads

➢ **20% average performance degradation** for Spark apps

➢ Not akin across all benchmarks

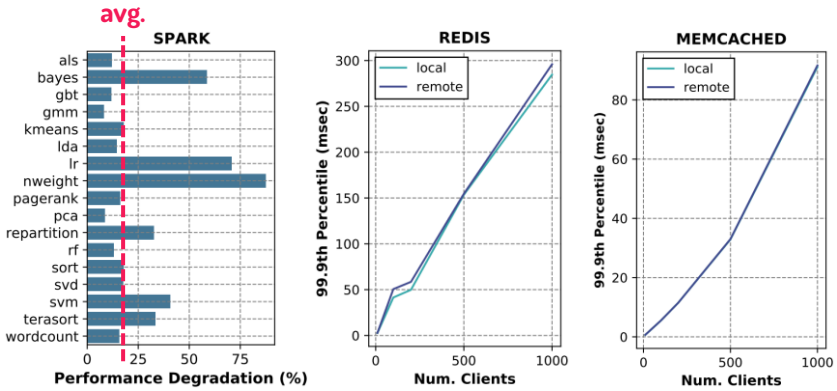
➢ **LC apps insensitive to remote memory**

Adrias: Interference-Aware Memory Orchestration for Disaggregated Cloud Infrastru

Dimosthenis Masouros
2023 IEEE International

ThymesisFlow: A Software-Defined, HW/SW co-Designed Interconnect Stack for Rack-Scale Memory Disaggregation

Christian Pinto, D. Syrivelis, +4 authors H. P. Hofstee • Published 1 October 2020 • Computer Science •
2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)



Performance Implications of Memory Disaggregation

- Performance degradation by design **(OR NOT?)** 🤔
  - Network Latency/Bandwidth & Protocol overheads

➢ **20% average performance degradation** for Spark apps

➢ Not akin across all benchmarks

➢ **LC apps insensitive to remote memory**

- Interference complicates things
  - Huge performance chasm between local and remote
  - Stacking interference effects

➢ **Up to x4 worst performance** under memory bandwidth and LLC interference

Adrias: Interference-Aware Memory Orchestration for Disaggregated Cloud Infrastru

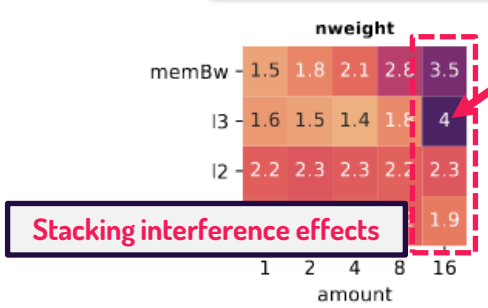
Dimosthenis Masouros

2023 IEEE International

ThymesisFlow: A Software-Defined, HW/SW co-Designed Interconnect Stack for Rack-Scale Memory Disaggregation

Christian Pinto, D. Syrivelis, +4 authors H. P. Hofstee • Published 1 October 2020 • Computer Science •

2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)



Up to x4 worst performance when stressing L3 cache

Stacking interference effects

Values show the slowdown of remote memory vs. local for the same amount of applied interference (**higher is worse**)

Performance Implications of Memory Disaggregation

- Performance degradation by design **(OR NOT?)** 🤔
  - Network Latency/Bandwidth & Protocol overheads

➢ **20% average performance degradation** for Spark apps

➢ Not akin across all benchmarks

➢ **LC apps insensitive to remote memory**

- Interference complicates things
  - Huge performance chasm between local and remote
  - Stacking interference effects

➢ **Up to x4 worst performance** under memory bandwidth and LLC interference

➢ Not akin across all benchmarks

Adrias: Interference-Aware Memory Orchestration for Disaggregated Cloud Infrastru

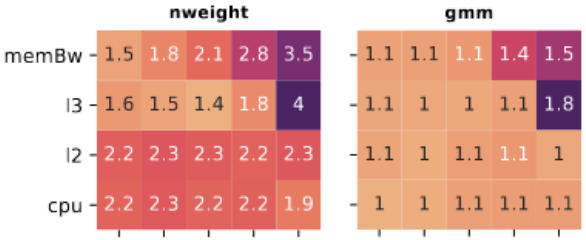
Dimosthenis Masouros

2023 IEEE International

ThymesisFlow: A Software-Defined, HW/SW co-Designed Interconnect Stack for Rack-Scale Memory Disaggregation

Christian Pinto, D. Syrivelis, +4 authors H. P. Hofstee • Published 1 October 2020 • Computer Science •

2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)



Not similar behavior across different benchmarks

Values show the slowdown of remote memory vs. local for the same amount of applied interference (**higher is worse**)

Performance Implications of Memory Disaggregation

- Performance degradation by design **(OR NOT?)** 🤔
  - Network Latency/Bandwidth & Protocol overheads

➢ **20% average performance degradation** for Spark apps

➢ Not akin across all benchmarks

➢ **LC apps insensitive to remote memory**

- Interference complicates things
  - Huge performance chasm between local and remote
  - Stacking interference effects

➢ **Up to x4 worst performance** under memory bandwidth and LLC interference

➢ Not akin across all benchmarks

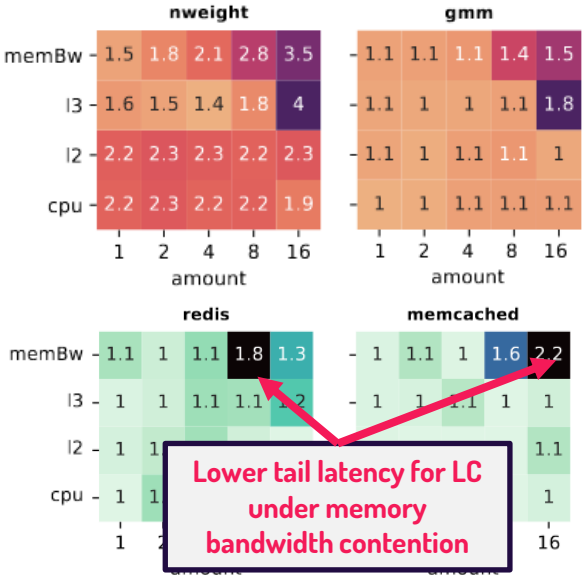
➢ **LC apps sensitive to memory bandwidth interference**

Adrias: Interference-Aware Memory Orchestration for Disaggregated Cloud Infrastru

Dimosthenis Masouros
2023 IEEE International

ThymesisFlow: A Software-Defined, HW/SW co-Designed Interconnect Stack for Rack-Scale Memory Disaggregation

Christian Pinto, D. Syrivelis, +4 authors H. P. Hofstee • Published 1 October 2020 • Computer Science •
2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)



Values show the slowdown of remote memory vs. local for the same amount of applied interference (**higher is worse**)

Potential Use of ML for Resource Management

- Naïve use of remote memory ➔ huge performance degradation
- "Intelligent" memory mapping of applications
  - Allocate "remote-memory friendly" apps on disaggregated pool
  - Minimize shared resource interference
  - Minimize data travelling back & forth through the network

Fundamental Questions

WHERE ?

HOW ?

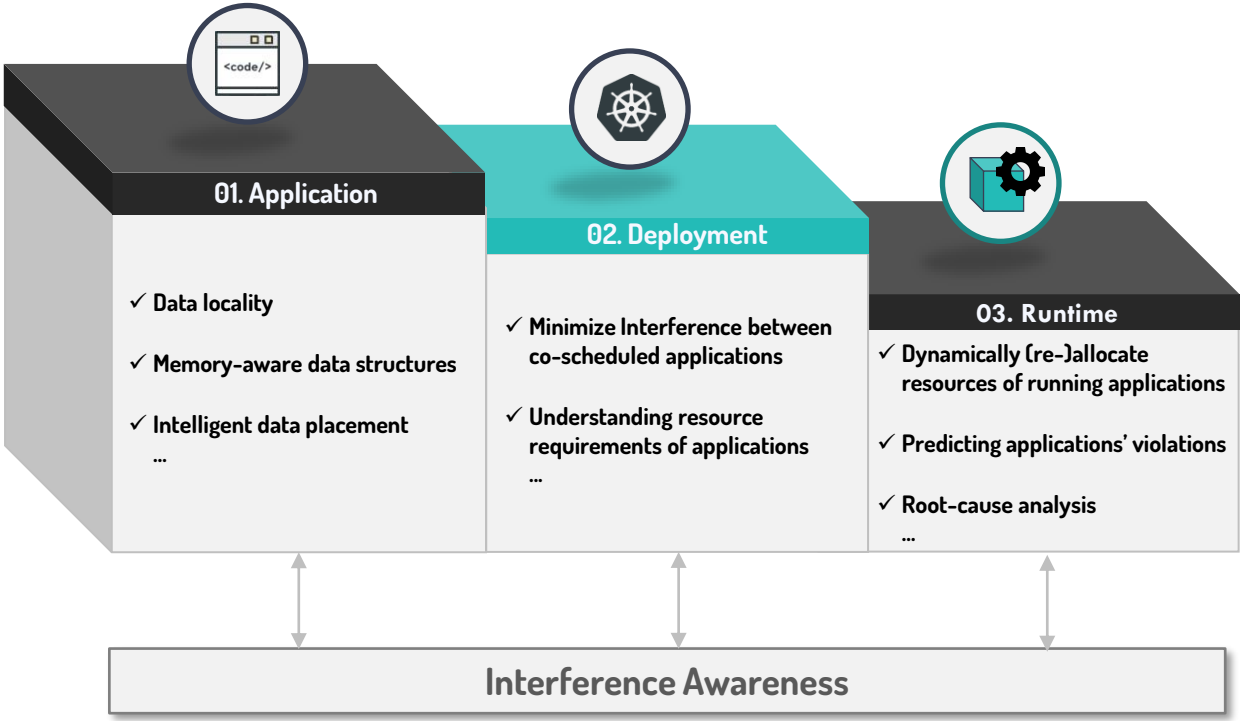
Fundamental Questions

WHERE ?

HOW ?

Levels of Applying ML Solutions

Software



Hardware

Prior works to be considered

Towards making the most of NLP-based device mapping optimization for OpenCL kernels
Petros Vavaroutsos, Ioannis Oroutzoglou, +1 author D. Soudris • Published 1 August 2022 • Computer Science •
2022 IEEE International Conference on Omni-layer Intelligent Systems (COINS)

End-to-End Deep Learning of Optimization Heuristics
Chris Cummins, Pavlos Petoumenos, +1 author H. Leather • Published 9 September 2017 • Computer Science •
Parallel Architectures and Compilation Techniques (PACT)

CodeBERT: A Pre-Trained Model for Programming and Natural Languages
Zhangyin Feng, Daya Guo, +8 authors Ming Zhou • Published 19 February 2020 • Computer Science • ArXiv

---

Sinan: ML-based and QoS-aware resource management for cloud microservices
Yanqi Zhang, Weizhe Hua, +3 authors WeizheHua • Published 19 April 2021 • Computer Science •

FIRM: An Intelligent Fine-Grained Resource Management Framework for SLO-Oriented Microservices
Subho Sankar Banerjee, +2 authors R. Iyer • Published in USENIX Symposium on Operating... 19 August 2020 •

Resource Central: Understanding and Predicting Workloads for Improved Resource Manageme in Large Cloud Platforms
Eli Cortez C. Vilarinho, Anand Bonde, +3 authors R. Bianchini • Published 14 October 2017 • Computer Science •
Proceedings of the 26th Symposium on Operating Systems Principles

Paragon: QoS-aware scheduling for heterogeneous datacenters
Christina Delimitrou, C. Kozyrakis • Published in International Conference on... 16 March 2013 • Computer Science

---

Learning Memory Access I
Milad Hashemi, Kevin Swersky, +5 authors Parthasarathy Ranganatha
Published in International Conference on... 6 March 2018 • Computer

Adjacent LSTM-Based Page Sch Hybrid DRAM/NVM Memory Sy
Manolis Katsaragakis, Konstantinos Stavrakakis, +2 authors D. Soudris • Published

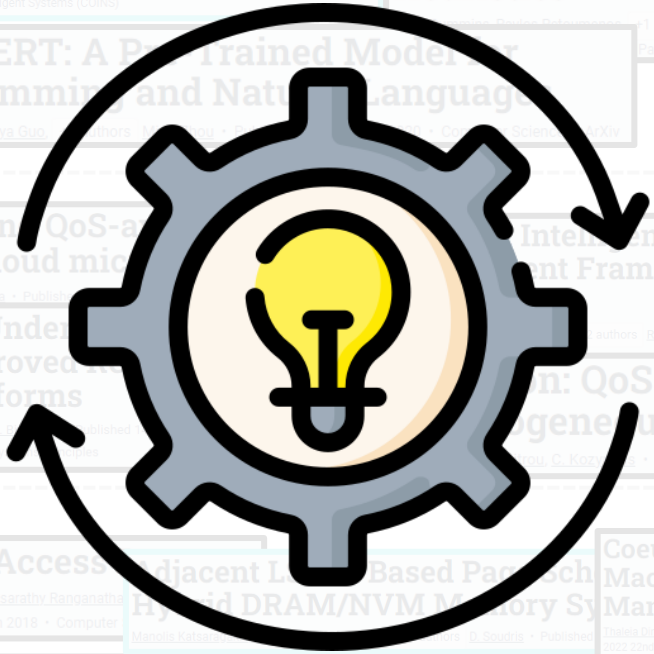
Coeus: Clustering (A)like Patterns for Practical Machine Intelligent Hybrid Memory Management
Thaleia Dimitra Doudali, Ada Gavrilovska • Published 1 May 2022 • Computer Science •
2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid)

Kleio: A Hybrid Memory Page S Machine Intelligence
Thaleia Dimitra Doudali, S. Blagodurov, +2 authors Ada Gavrilovska • Published 1
Proceedings of the 28th International Symposium on High-Performance Parallel and

Nimble Page Management for Tiered Memory Systems
Zi Yan, Daniel Lustig, +1 author A. Bhattacharjee • Published 4 April
Proceedings of the Twenty-Fourth International Conference on Architec
Systems

Utility-Based Hybrid Memory Management
Y. Li, Saugata Ghose, +3 authors O. Mutlu • Published 1 September 2017 • Computer Science •
2017 IEEE International Conference on Cluster Computing (CLUSTER)

Prior works to be considered



Fundamental Questions

WHERE ?

HOW ?

How to use ML?

Things to consider when deciding how to exploit ML in cloud resource management:

- What should the ML model do?
- What inputs should the ML model receive?
- What should the ML model's architecture be?
- How quickly and often should the ML do predictions?

…

Exploring the opportunities to use ML, the possible designs, and our experience with Microsoft Azure.

BY RICARDO BIANCHINI, MARCUS FONTOURA, ELI CORTEZ, ANAND BONDE, ALEXANDRE MUZIO, ANA-MARIA CONSTANTIN, THOMAS MOSCIBRODA, GABRIEL MAGALHAES, GIRISH BABLANI, AND MARK RUSSINOVICH

Toward ML-Centric Cloud Platforms

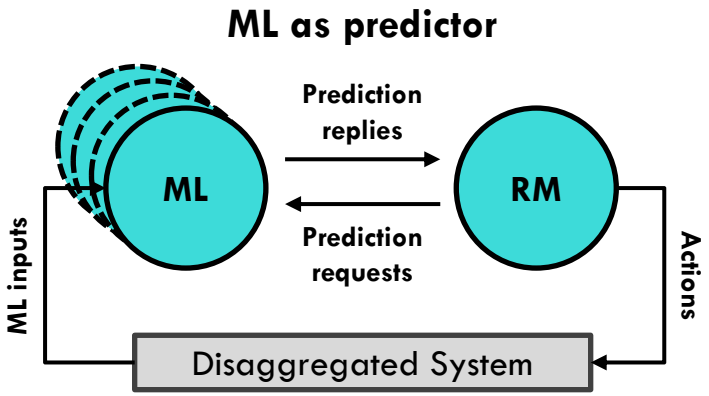
ML Integration Approaches

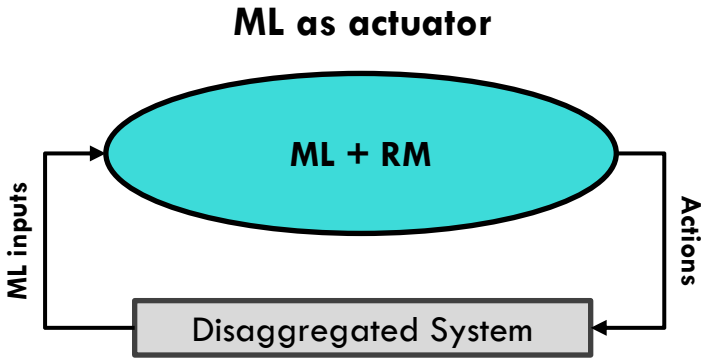
ML model's purpose:

- Behavioral predictions (predictor)
- Actual management actions (actuator)

ML modeling approach:

- One model per application / memory type
  + More accurate (probably)
  - Less scalable
- One model to rule them all!





What Inputs to Use?

- Application-specific (high-level) metrics
  - Accurate (real) performance
  - Difficult to acquire
  - Developers have to instrument apps with monitoring tools (e.g., AWS Cloudwatch)
- System-wide (lower-level) metrics
  - Imperfect proxy for performance
  - Always available to providers
  - Can be used to "predict" performance

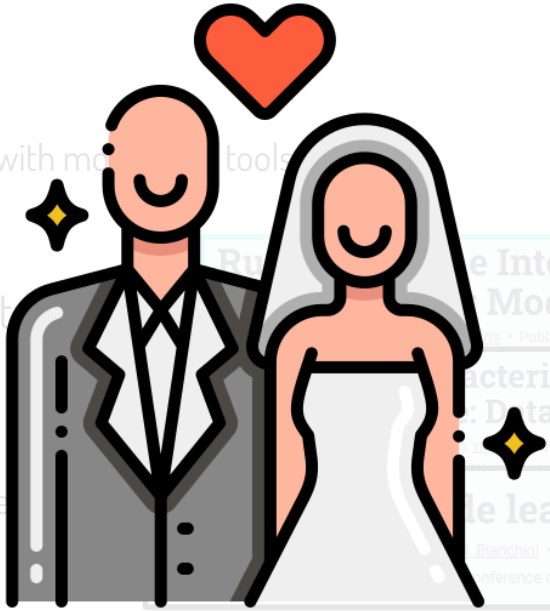
Rusty: Runtime Interference-Aware Predictive Monitoring for Modern Multi-Tenant Systems
Dimosthenis Masouros, S. Xydis, D. Soudris • Published 1 January 2021 • Computer Science •
IEEE Transactions on

Characterizing Job Microarchitectural Profiles at Scale: Dataset and Analysis
Kangjin Wang, Ying Li, +8 authors Liping Zhang • Published 29 August 2022 • Computer Science •

SOL: safe on-node learning in cloud platforms
Ya-wen Wang, D. Crankshaw, +3 authors R. Bianchini • Published 25 January 2022 • Computer Science •
Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems

What Inputs to Use?

- Application-specific (high-level) metrics
  - Accurate (real) performance
  - Difficult to acquire
  - Developers have to instrument apps with mo tools (e.g., AWS Cloudwatch)
- System-wide (lower-level) met
  - Imperfect proxy for performance
  - Always available to providers
  - Can be used to "predict" performance



e Interference-Aware Predictive
Modern Multi-Tenant Systems
s • Published 1 January 2021 • Computer Science •

acterizing Job Microarchitectural Profiles at
Dataset and Analysis
uthors Liping Zhang • Published 29 August 2022 • Computer Science •

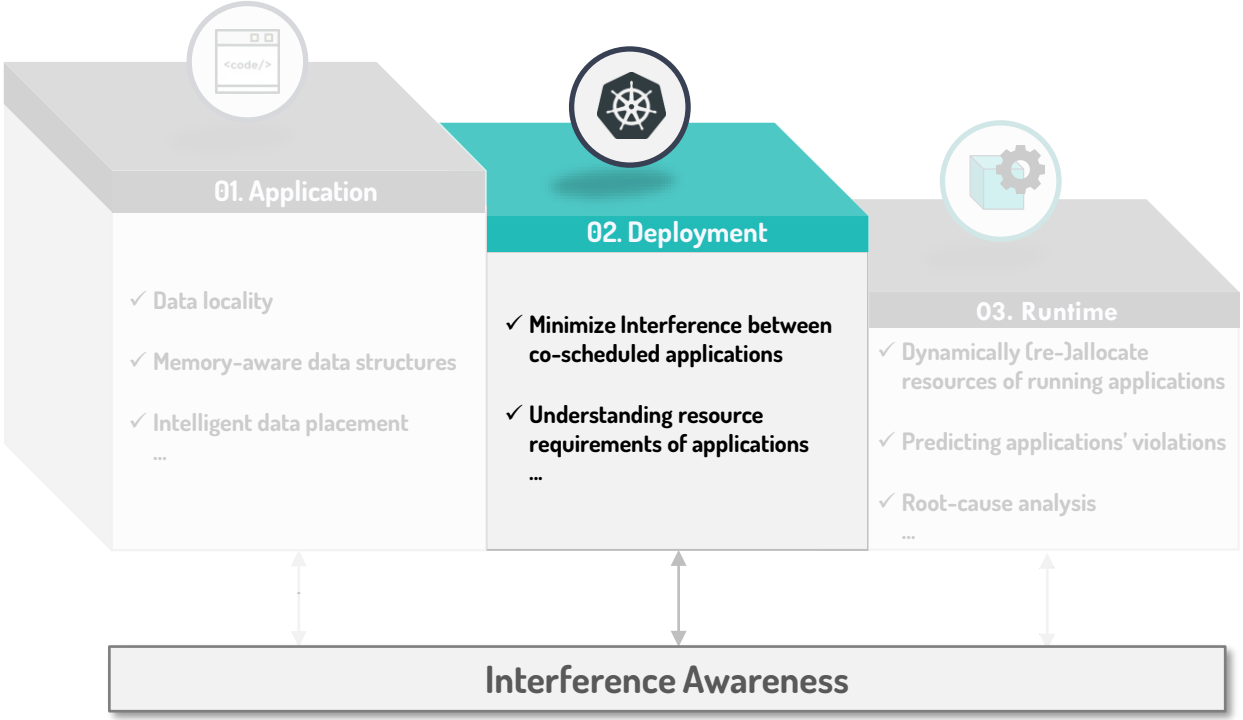
le learning in cloud platforms
Bianchini • Published 25 January 2022 • Computer Science •
nference on Architectural Support for Programming Languages and Operating Systems

Adrias

Interference-Aware Memory Orchestration for Disaggregated Cloud Infrastructures

Levels of Applying ML Solutions

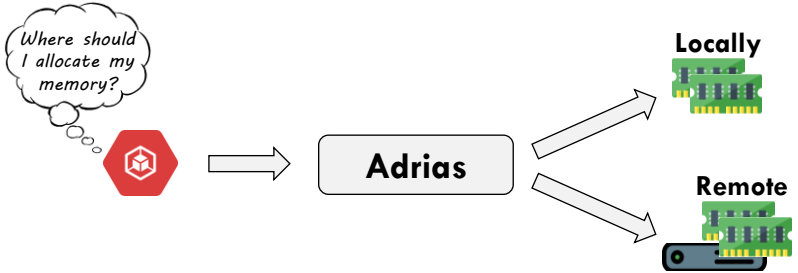
Software

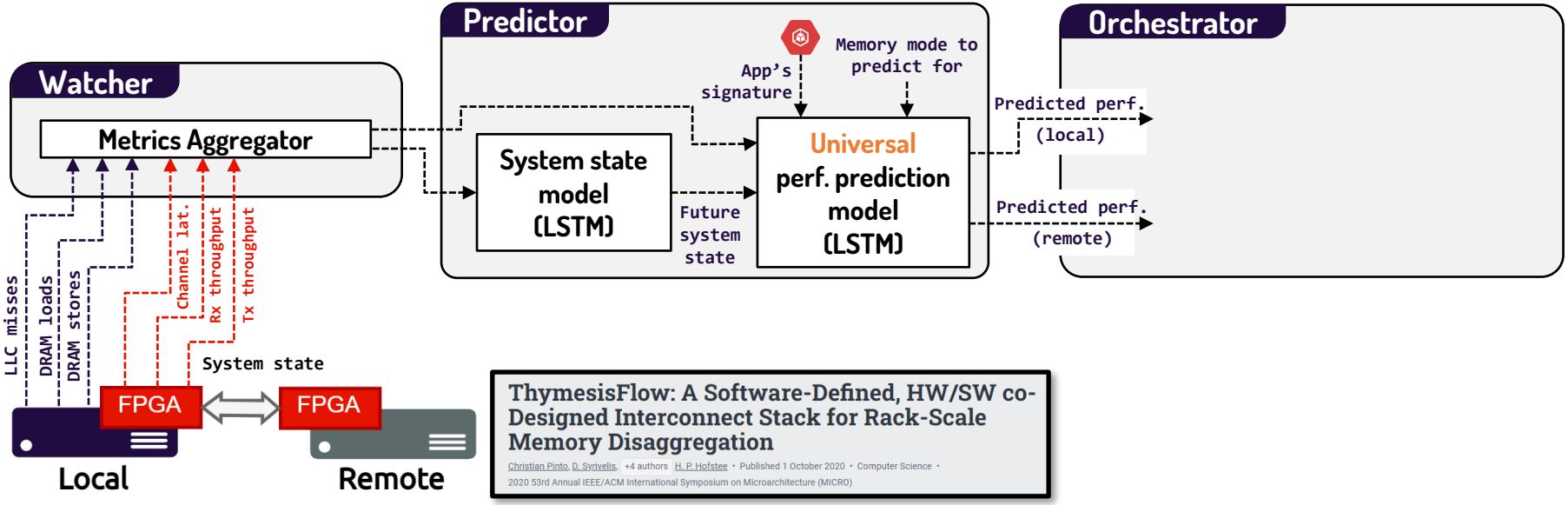


Hardware

(really abstract) Overview

Goal: Adrias aims to efficiently **place all memory allocations of incoming applications either on local or remote memory**



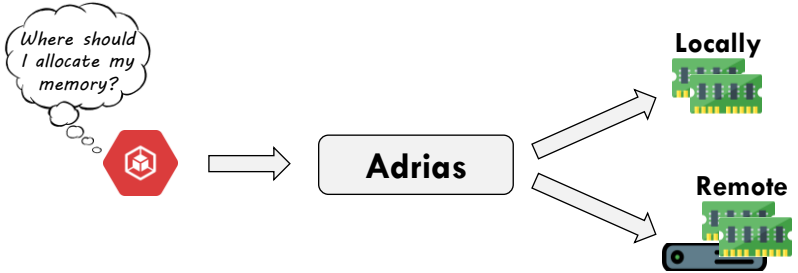


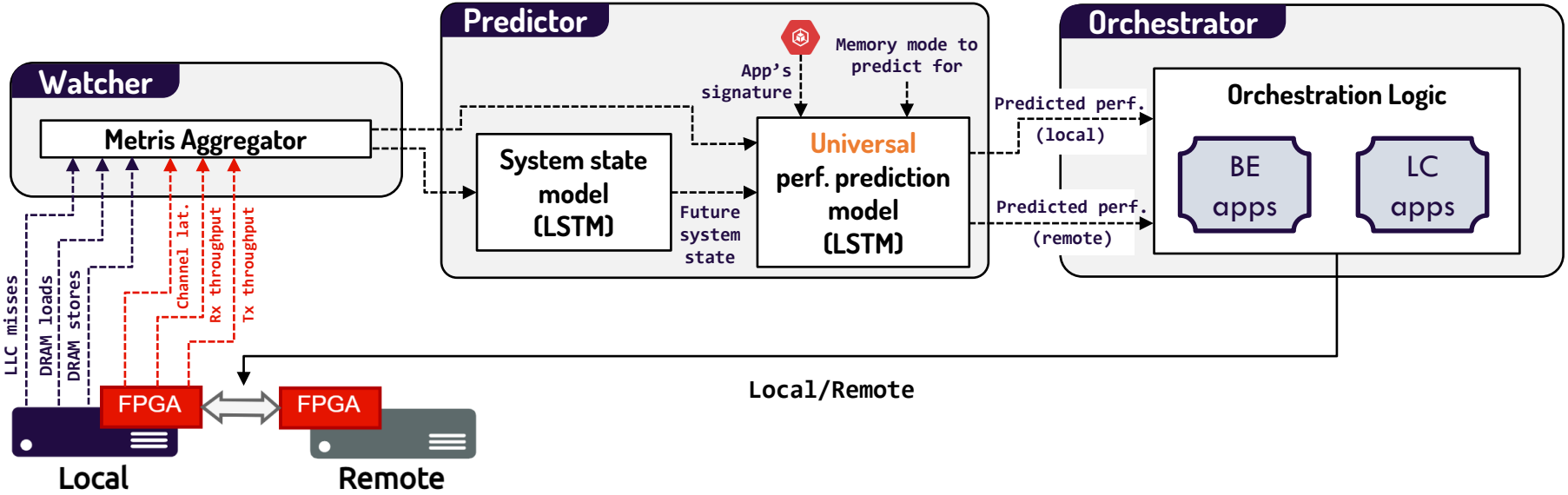
ThymesisFlow: A Software-Defined, HW/SW co-Designed Interconnect Stack for Rack-Scale Memory Disaggregation

Christian Pinto, D. Syrivelis, +4 authors H. P. Hofstee • Published 1 October 2020 • Computer Science •
2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)

(really abstract) Overview

Goal: Adrias aims to efficiently **place all memory allocations of incoming applications either on local or remote memory**



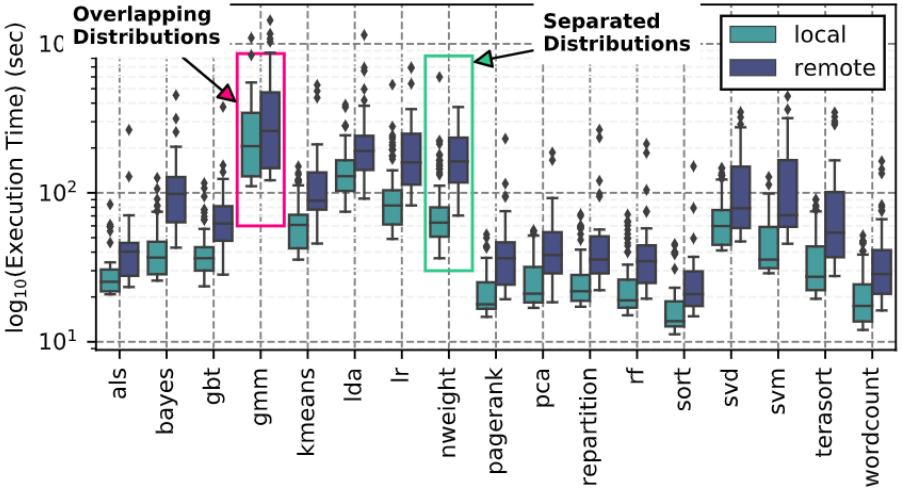


Data collection

- 72 random 1-hour scenarios deployed
  - Random applications arriving at random intervals

Scenarios' Insights

- **Overlapping performance distributions** (e.g., gmm)
  - Remote memory could be beneficial in certain interference scenarios
  - Sacrifice performance for leveraging remote memory
- **Separated performance distributions** (e.g., nweight)
  - Use of remote memory prohibitive due to stacking interference effects

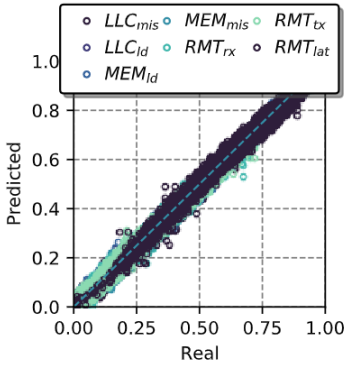


Evaluation Results

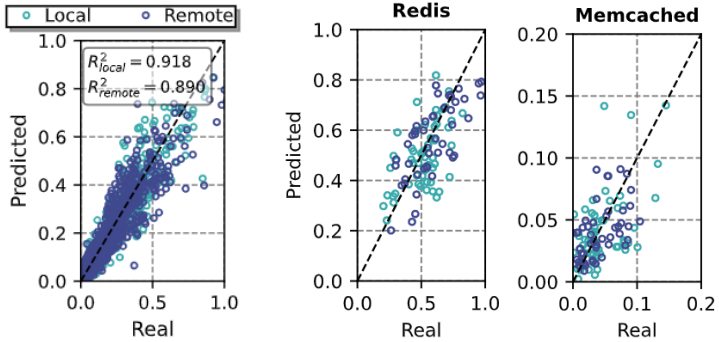
- System state model
  - **~0.99** R^2 score* for all predicted metrics
- Performance prediction model
  - **0.94 average** R^2 score* for performance prediction of both BE and LC on local and remote
- Orchestration logic
  - Outperforms Random and Round-Robin schedulers
  - Allocates 10%-35% to remote memory with 0.5%-15% median performance degradation
  - Up to 55% less traffic on the network channel
  - Aligns with scenarios' insights

*Values closer to 1 are better

Predicted vs. Real (system state model)



Predicted vs. Real (perf. prediction model)



For More Results…

Adrias: Interference-Aware Memory Orchestration for Disaggregated Cloud Infrastructures

Dimosthenis Masouros, Christian Pinto, +2 authors D. Soudris • Published 1 February 2023 • Computer Science •
2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)

Conclusions

- Memory disaggregation is the next big thing (in the Cloud)
- Memory disaggregation + interference ➔ huge performance variability and unpredictability
- ML for systems can be a strong tool (tailored to our needs)
- Rethink and adapt current ML solutions

Adrias showed the potential of ML in disaggregated memory systems

- Separated ML and Orchestration logic
- Able to leverage remote memory with minimal performance impact

Q & A

dmasouros@microlab.ntua.gr