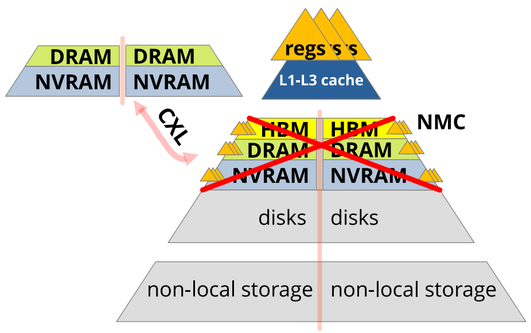
SMAUG: System-Level Modeling and Optimized Use of Disruptive Memory Technologies

UNIVERSITÄT OSNABRÜCK

embedded software systems

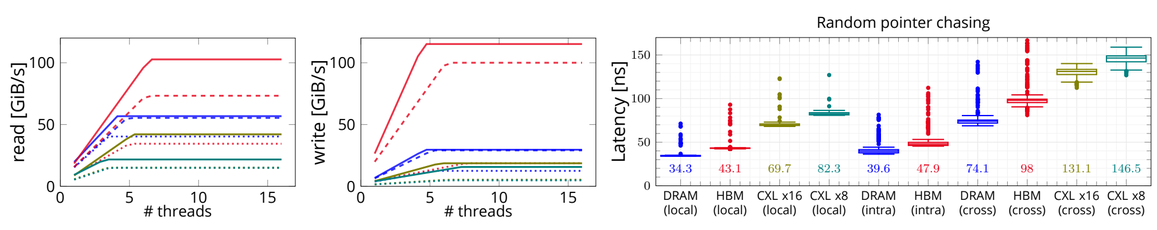
Motivation: Complex Memory Landscape



DMTs break out of the memory hierarchy: conventional placement and optimization algorithms are no longer valid; state of the art is often limited to offline strategies and single DMTs

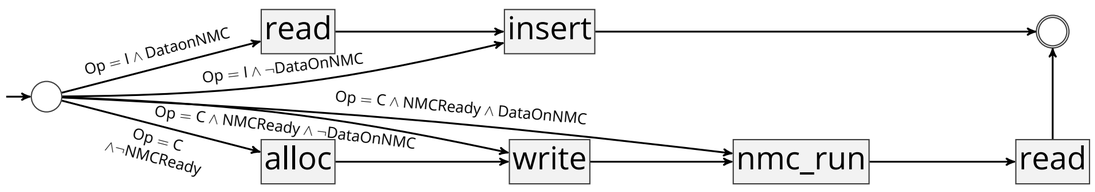
⇒ SMAUG: **whole-system** models and algorithms
⇒ Optimal placement of compute and data

Platform Models for CXL and HBM



- local CXL ≈ remote DRAM (latency, throughput)
- HBM: latency can **slow down** applications [FLS24]

Application Models for Runtime Behaviour



$23.3\,\text{ms}$	$4.80\,\frac{\text{GB}}{\text{s}} + 0.35\,\frac{\text{GB}}{\text{s}}$	$213\,\mu\text{s}$	$2.66\,\frac{\text{GB}}{\text{s}} + 0.17\,\frac{\text{GB}}{\text{s}}$
$+\ 2.5\,\text{ms} \cdot \#ranks$	$\cdot\min(\#ranks, 23)$	$+0.54\,\text{ns} \cdot \frac{\#rows}{\#ranks}$	$\cdot\min(\#ranks, 28)$

$\mathcal{A} = (Q, q_0, q_f, \Delta, P, \gamma, \rho, \lambda)$ — States ($\widehat{=}$ API calls) Q, transitions $\Delta \subseteq Q^2$, workload config domain P, transition guards $\gamma : \Delta \rightharpoonup (P \to \{\bot, \top\})$, loop count predictors $\rho : Q \to (P \to \mathbb{N})$, model mapping $\lambda : Q \to (P \to \mathbb{R}^{n_q})$

① Workload specification $P \mapsto$ API call sequences & args (γ, ρ: featured automata + regression model trees)

② For each API call: args $\mapsto$ performance attributes (λ: regression model trees)

$\to$ **Performance-Aware Behaviour Models** [FS25b]

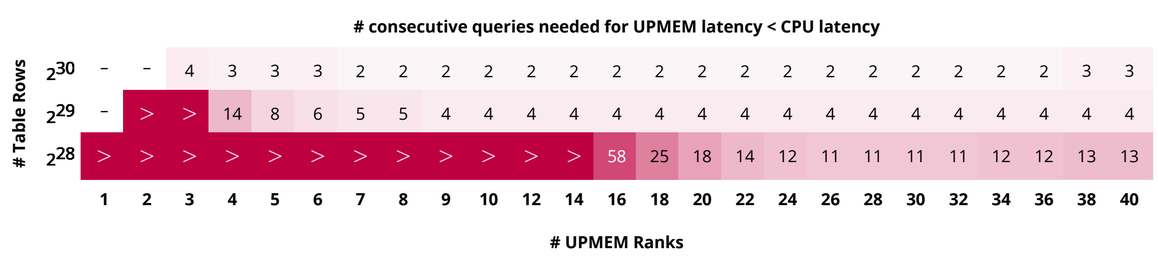
Learning:

① Application traces (AspectC++, simulator) $\to Q, \Delta, \gamma, \rho$

② Traces on real hardware or microbenchmarks $\to \lambda$

- Proof-of-concept with **UPMEM PIM SDK** [FS25a]

Key Findings



- UPMEM PIM: DBMS offloading only sensible for large columns and ≥ 2 **consecutive operations** [FLS25]
- Models **explain performance bottlenecks** [FS25c] (state of the art: runtime behaviour as black box)

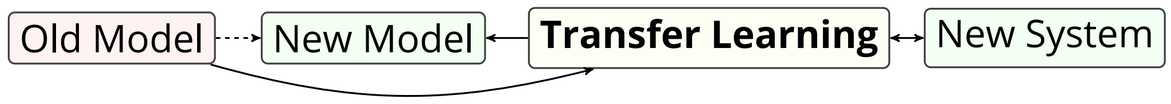
Second Funding Period

Goal: Use models from first period to provide **holistic, model-guided resource management strategies**.
⇒ Benefit of such strategies for systems with DMTs?
Research questions / corresponding research objectives:

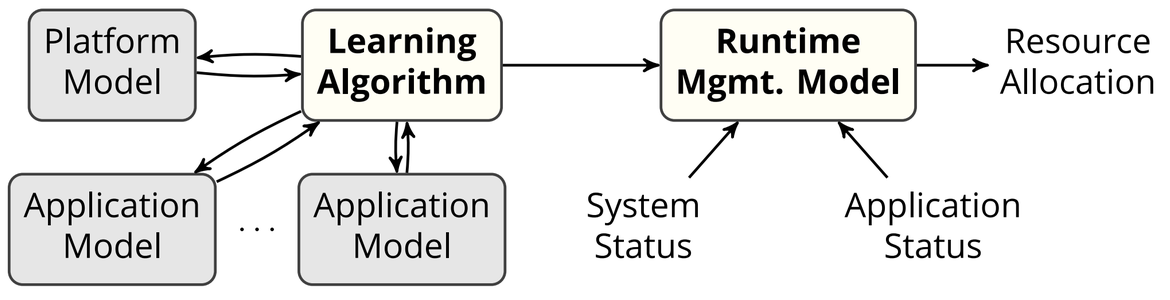
① System-wide resource partitioning strategies: coarse-grained partitioning to avoid interference

② Model-guided application performance optimization: fine-grained resource management within partitions

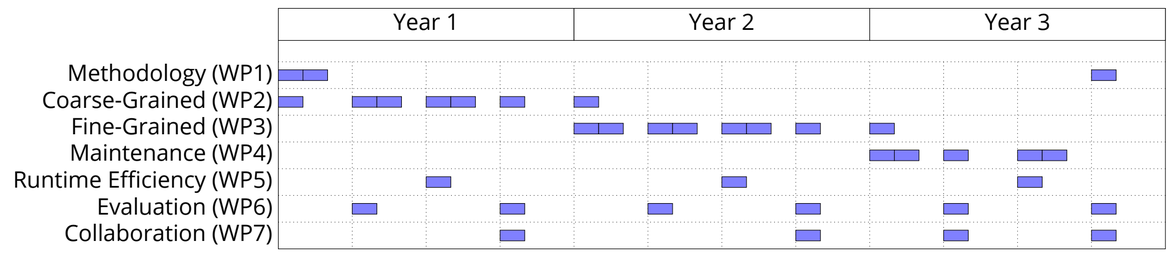
③ Model maintenance in practice: avoid re-training models from scratch after platform (HW/SW) changes



④ Efficient runtime queries: avoid slowing down time-critical decisions in resource management algorithms



Work Packages



- WP1: Encoding input data and decisions for RMTs
- WP2 ... WP5 $\widehat{=}$ research questions ① ... ④
- WP6: HetSim (simulator) + MxKernel (research OS)
- WP7: HetCIM, Memento, PIMPMe, VAMPIR, *you?*

References

[FLS24] Birte Friesel, Marcel Lütke Dreimann, and Olaf Spinczyk. "Performance Models for Task-based Scheduling with Disruptive Memory Technologies". In: *Proc. 2nd Workshop on Disruptive Memory Systems*. DIMES '24. Austin, TX, USA: ACM, Nov. 2024, pp. 1–8. DOI: 10.1145/3698783.3699376.

[FLS25] Birte Friesel, Marcel Lütke Dreimann, and Olaf Spinczyk. "Lightning Talk: Feasibility Analysis of Semi-Permanent Database Offloading to UPMEM Near-Memory Computing Modules". In: *Datenbanksysteme für Business, Technologie und Web – Workshopband*. BTW '25. Bamberg, Germany: GI, Mar. 2025, pp. 355–366. DOI: 10.18420/BTW2025-140.

[FS25a] Birte Friesel and Olaf Spinczyk. "Overhead Prediction for PIM-Enabled Applications with Performance-Aware Behaviour Models". In: *Proc. 1st Cross-disciplinary Conference on Memory-Centric Computing*. CCMCC '25. to appear. Dresden, Germany: IEEE, Oct. 2025.

[FS25b] Birte Friesel and Olaf Spinczyk. "Performance-Aware Behaviour Models for Feature-Dependent Runtime Attributes in Product Lines". In: *Proc. 19th International Working Conference on Variability Modelling of Software-Intensive Systems*. VaMoS '25. Rennes, France: ACM, Feb. 2025, pp. 131–135. DOI: 10.1145/3715340.3715435.

[FS25c] Birte Friesel and Olaf Spinczyk. "Understanding Product Line Runtime Performance with Behaviour Models and Regression Model Trees". In: *Proc. 29th International Systems and Software Product Line Conference*. SPLC-A '25. A Coruña, Spain: ACM, Sept. 2025. DOI: 10.1145/3744915.3748472.

Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation

Birte Friesel (PI), **Michael Müller** and **Olaf Spinczyk** (PI)
{birte.friesel,michael.mueller,olaf}@uos.de · ess.cs.uos.de

SPP 2377