



OMI in a Composable World - A Software Perspective

Felix Eberhardt, Andreas Grapentin, Sven Köhler, Lukas Wenzel, Prof. Andreas Polze

firstname.lastname@hpi.de

Operating Systems and Middleware Group,

Hasso Plattner Institute

02.08.2022



Research Areas


- Small / embedded systems / IoT
 - Real-time, Fault-Tolerance, Security
 - Telemed5000 Project – Charité
 - Rail2X, RailChain, DiAK – Deutsche Bahn
- Big / server systems / data center
 - Energy-aware computing
 - Accelerators (FGPAs, GPUs, ...)
 - Memory Subsystem (disaggregation, ...)



**OMI - Software
Perspective**

OSM Group

Two Trends:

- Hardware (managed by OS) increasingly **Complex** and **Heterogeneous**
 - Software (managed by OS) increasingly **Chaotic** and **Resource-Hungry**
- 

Fog of War

- Complexity and Specifics of the Hardware are **hidden** from Applications through Programming Models and Abstractions (this is a good thing)
- OS as all-knowing benevolent entity manages access to resources
- Some abstractions exist to query hardware characteristics:
 - libNUMA API, Perf Counters, proc pseudofilesystem kernel breadcrumbs

**OMI - Software
Perspective**

OSM Group

BUT:

- Applications (and Middlewares) need cues from Hardware and OS to make informed decisions at runtime regarding many issues, such as:
 - Long-term Scheduling decisions
 - Memory allocations to match functional / non-functional constraints
- While many probes exist, they are **spread** across various API, **difficult** to integrate into existing applications, and not well **documented**.
- Consequently, We are not as good at efficiently utilizing Heterogeneous system resources and Energy-aware computing as we should be.

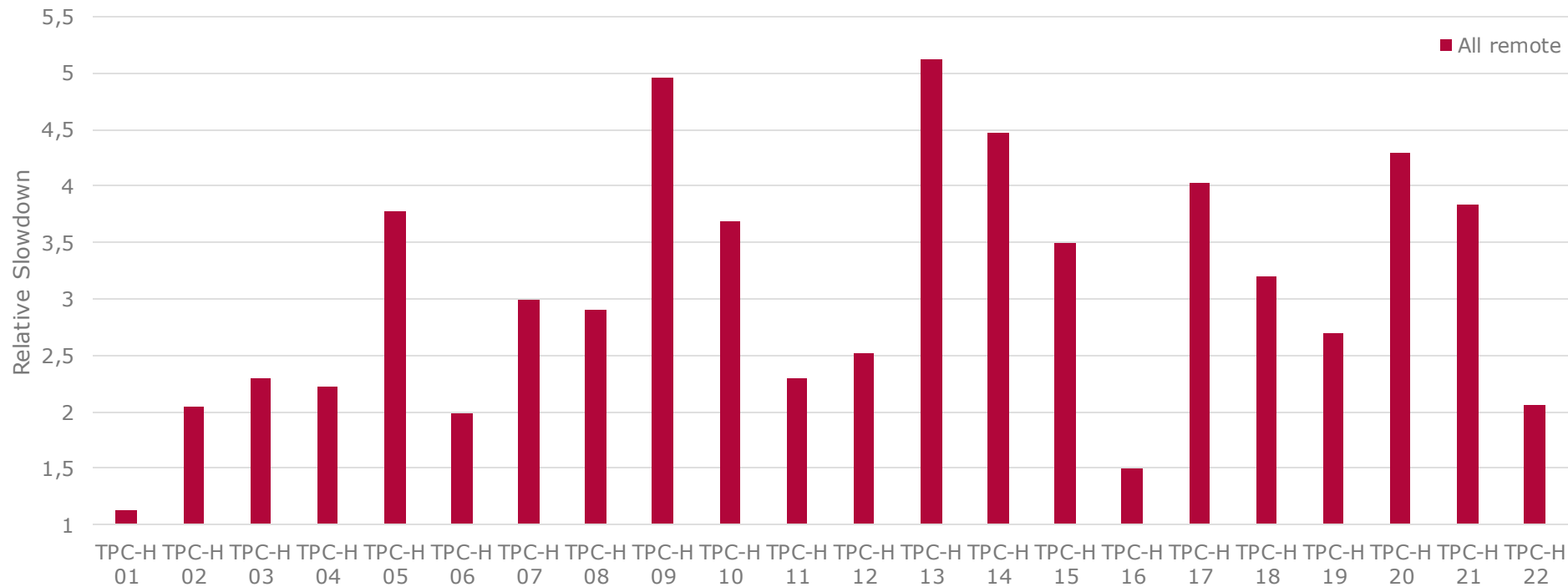
**OMI - Software
Perspective**

OSM Group

We need support throughout the full stack of HW, OS, MW, APP

Paying the Price of Heterogeneous Latency

TPC-H on local vs. disaggregated Memory



■ SF 10 ~ 10 GB

■ Some queries more severe impacted than others

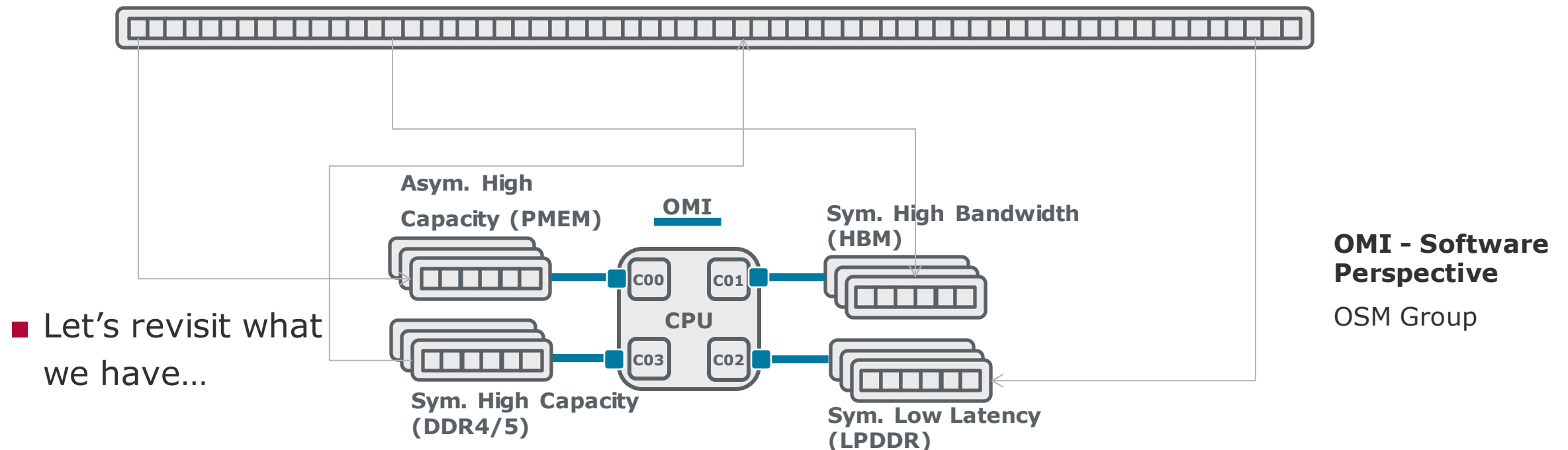
■ Some queries have nearly no impact

**OMI - Software
Perspective**

OSM Group

OMI Enablement Heterogeneous Memories!

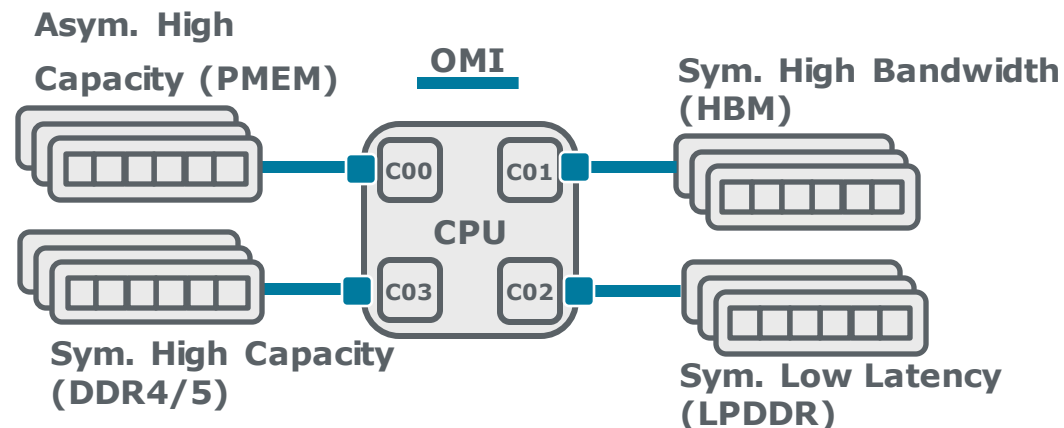
- OMI enables domain specific memory types
- Have different **functional** and **non-functional** characteristics, most notably different energy characteristics
- Managed by OS or middleware or application -- Transparent vs explicit API



OMI Enablement

What about the OS?

- two OS abstraction we explored: NUMA node (Scale-Up machines)... or memory mapped files (MULTICS)
- But: even NUMA-aware applications are not fully prepared for vastly different performance characteristics
- Intel PMEM made applications aware, but too vendor specific
- Energy-awareness is not addressed
- Disaggregation to increase resource usage...



```
bash-4.4$ numactl -H
available: 4 nodes (0,8,16,32)
node 0 cpus: 0 1 2 3 4 5 6 7 8 9 10
5 46 47 48 49 50 51 52 53 54 55 56 5
node 0 size: 261649 MB
node 0 free: 260828 MB
node 8 cpus: 64 65 66 67 68 69 70 71
4 105 106 107 108 109 110 111 112 11
node 8 size: 258659 MB
node 8 free: 253826 MB
node 16 cpus:
node 16 size: 512 MB
node 16 free: 512 MB
node 32 cpus:
node 32 size: 0 MB
node 32 free: 0 MB
node distances:
node 0 8 16 32
0: 10 40 80 80
8: 40 10 80 80
16: 80 80 10 80
32: 80 80 80 10
```

**OMI - Software
Perspective**

OSM Group

Disaggregated Memory Testbed

Thymesisflow as Heterogeneity Superlative

■ Hardware

- 2x IBM **IC922** or IBM AC922 or Inspur FP5290G2
- 2x AlphaData 9V3 FPGAs
- 2x OpenCAPI cables (SlimSAS) 25 Gb/s x8
- 2x 100 Gb/s network cabling

Static Offsets

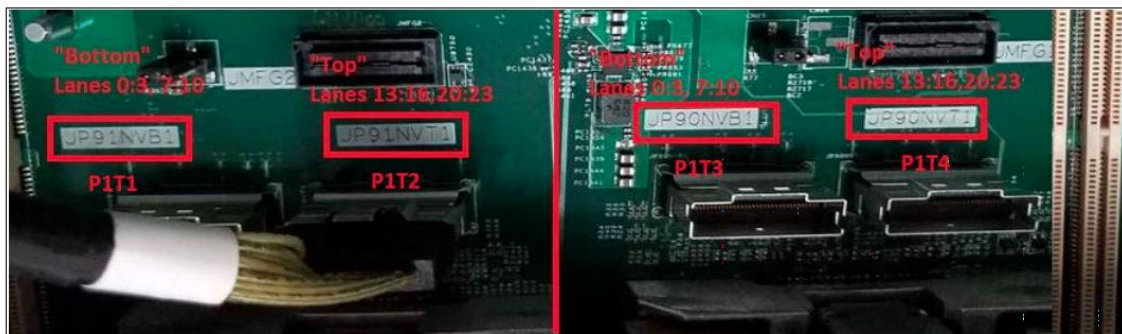
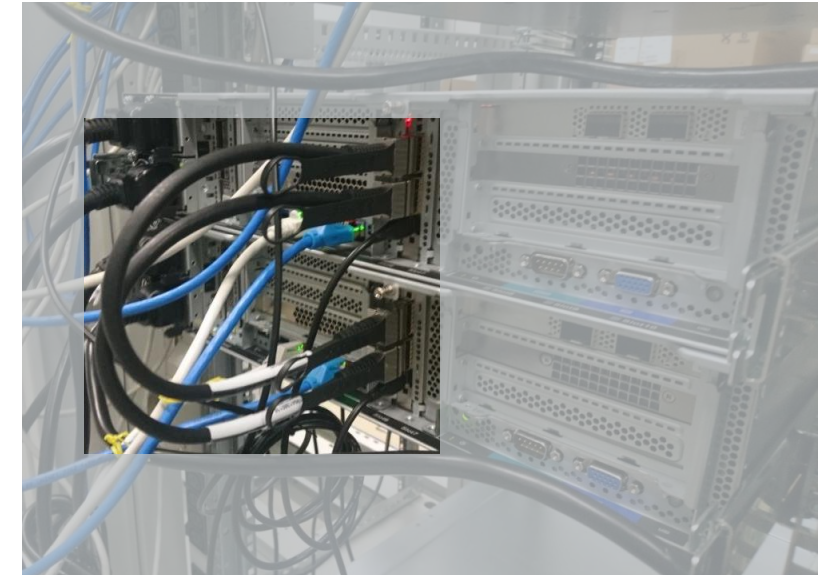


Figure 1-9 OpenCAPI 3.0 ports in the Power IC922 server

Processor 0:
0x2000000000000

Processor 1:
0x2200000000000

Back-to-back or cross?



```
// Ports specified as bitfield  
// Send Port  
rc = ocxl_mmio_write64(conn->global, 0x78, OCXL_MMIO_LITTLE_ENDIAN, 0x1);
```

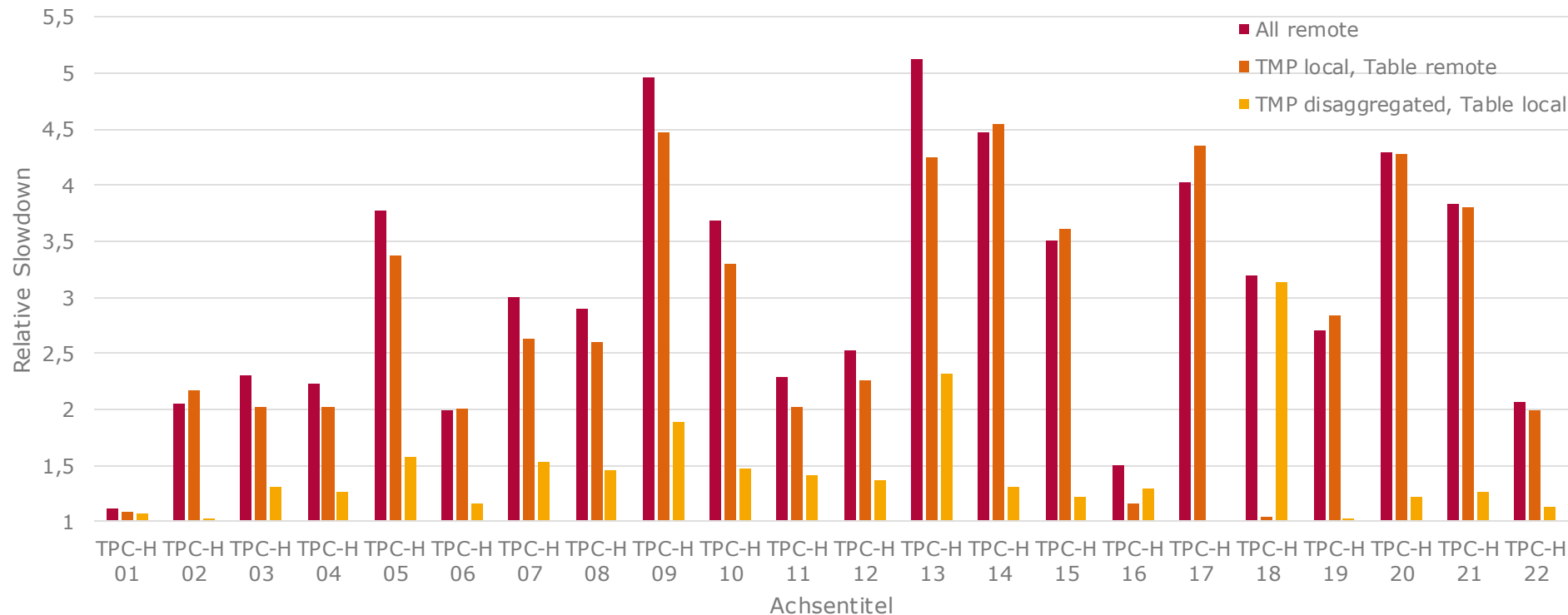
```
// Receive Port  
#define AFU_PORT 2
```

**OMI - Software
Perspective**

OSM Group

Disaggregated In-Memory Databases

TPC-H on local vs. disaggregated Memory

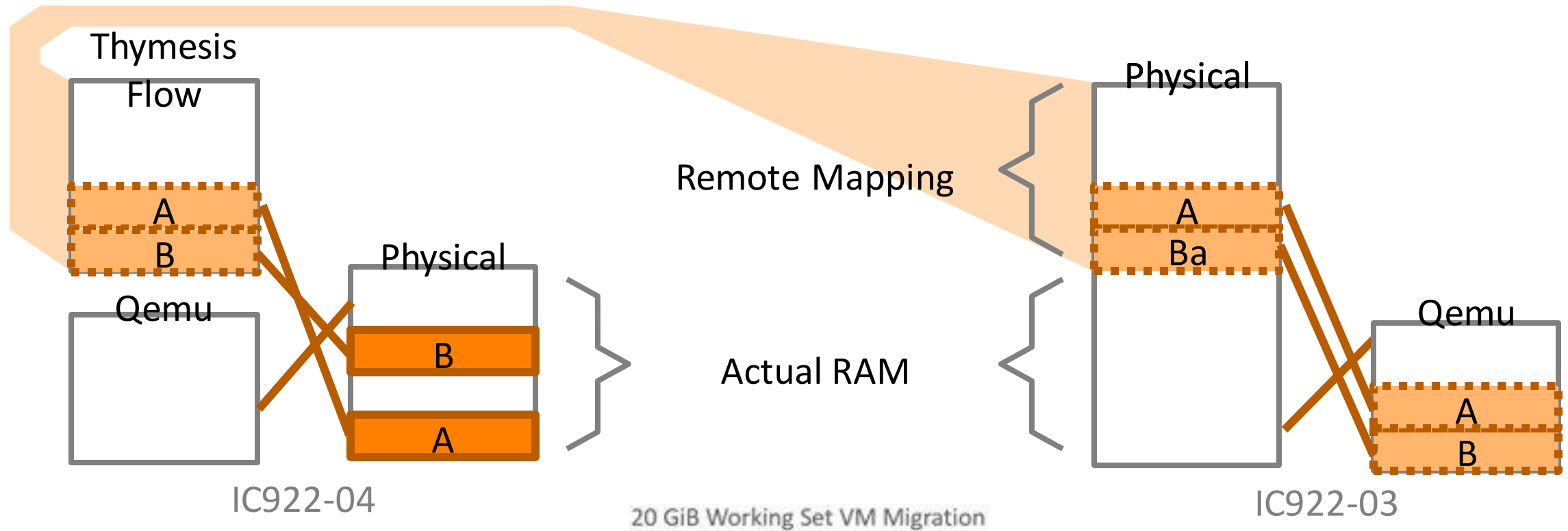


- SF 10 ~ 10 GB
- Needs modification of application
- Or regular migratepages to local node

OMI - Software Perspective

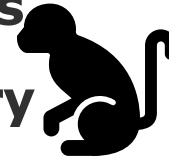
OSM Group

Zero-Copy VM Migration pushing the boundaries of virtualization



Two Trends:

- Hardware (managed by OS) increasingly **Complex** and **Heterogeneous**
- Software (managed by OS) increasingly **Chaotic** and **Resource-Hungry**
- **With growing Heterogeneity we pay the Price of Abstractions**
 - Energy Efficiency, Utilization, Performance, Throughput
- **Smarter Probes and Programming Abstractions are needed to consolidate the Perspective of Software on Future OMI-Enabled Heavily Heterogeneous Computing Systems**

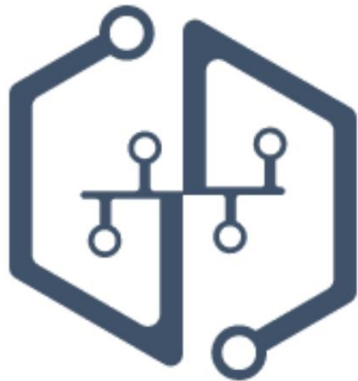


OMI - Software
Perspective
OSM Group

Invitation for Collaboration / Contribution Workshop / Lab Resources

<https://compsysworkshop.github.io/compsys22/>

COMPSYS '22



Workshop on Composable Systems
Co-located with [IPDPS 2022](#)

Date

- Workshop: May 30 through June 3 2022
- Second iteration planned for 2023

First Workshop on Composable Systems (COMPSYS '22)

Call for Papers

Papers are solicited from the areas, including, but not limited to:

- Hardware and emerging storage technologies
 - Hardware architectures for composability
 - Power, energy, and thermal management for composable systems
 - Memory and storage technologies for composable system
- Modeling, Prototyping and Evaluation
 - Composable system prototypes
 - Modeling of composable systems
 - Evaluation of applications on composable systems
 - Failure and resilience models for composable systems
- System software and programming models/tools
 - Control plane software for management of composable systems
 - Programming models for composable systems
 - Analysis / profiling tools and techniques for composable systems
 - Software runtimes for composability in Cloud and HPC
 - Virtualization for composable systems

are



OMI in a Composable World - A Software Perspective

Felix Eberhardt, Andreas Grapentin, Sven Köhler, Lukas Wenzel, Prof. Andreas Polze

firstname.lastname@hpi.de

Operating Systems and Middleware Group,

Hasso Plattner Institute

02.08.2022