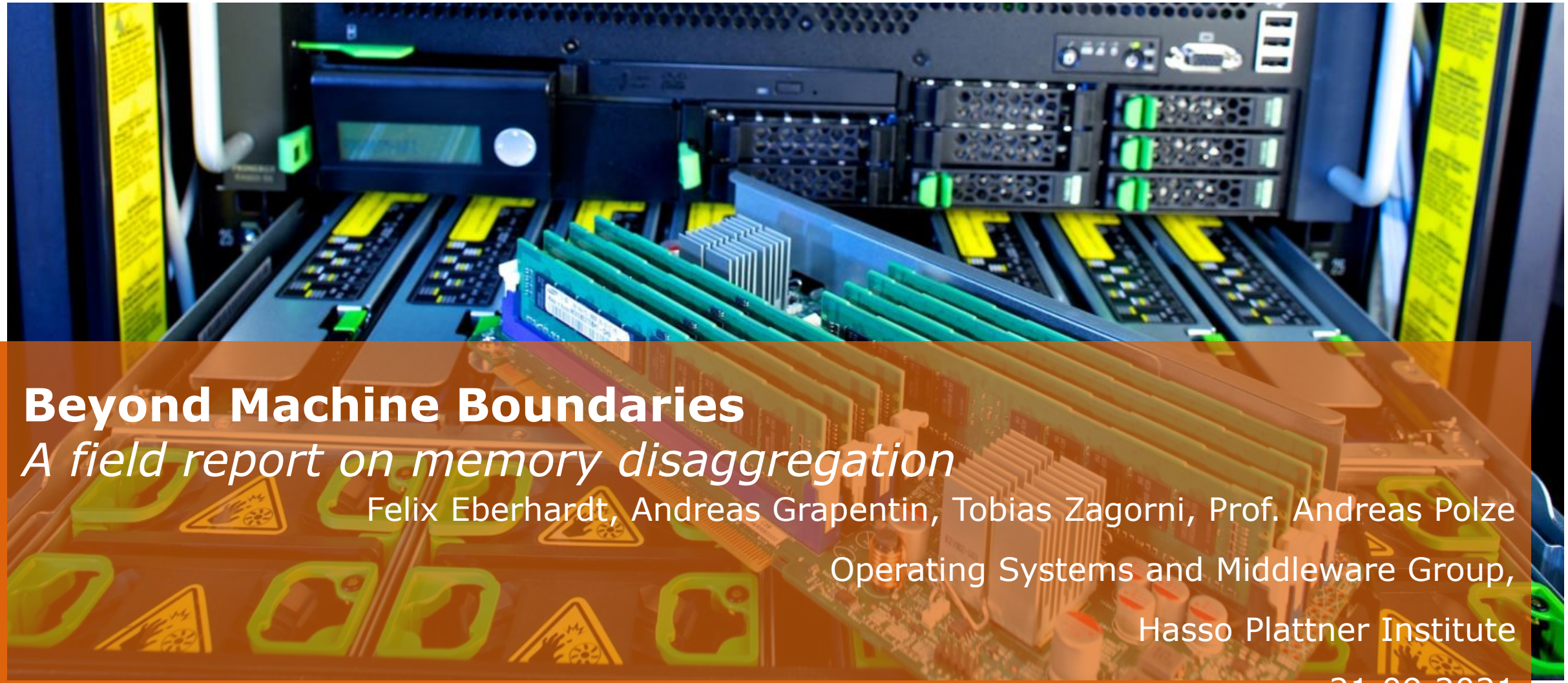
HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam



Beyond Machine Boundaries

A field report on memory disaggregation

Felix Eberhardt, Andreas Grapentin, Tobias Zagorni, Prof. Andreas Polze

Operating Systems and Middleware Group,

Hasso Plattner Institute

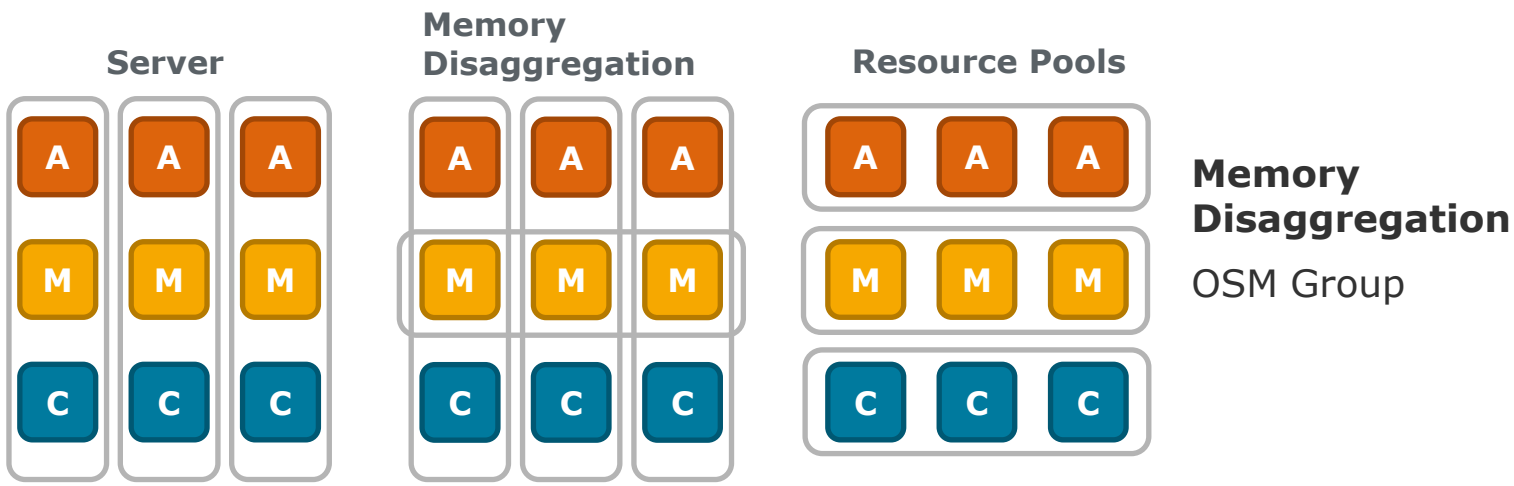
21.09.2021

Agenda

- Motivation & Coherent Interconnects
- ThymesisFlow Prototype
- Research Questions / Areas
- Projects
- Workshop IPDPS / Collaboration

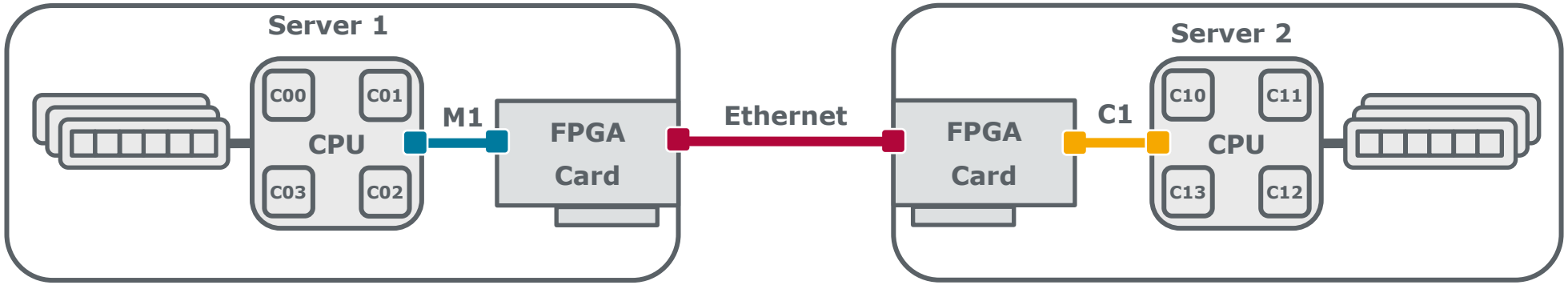
Resource Disaggregation
Motivation

- Workloads with different needs
  - With phases / different parts
  - How to map on hardware / virtual machine?
- With disaggregation more flexibility of deployment
  - Better utilization, fewer stranded resources
- Announced as product
  - Power 10 Memory Inception



Memory Disaggregation Technology

- Coherent Interconnects
  - Push (copy) vs. pull (coherent access)
  - No need to copy data but share virtual addresses
  - Several approaches: CXL, GenZ, OpenCAPI
- Mode of integration
  - Memory Mode (Memory controller): Cxl.memory or OpenCAPI M1
  - Compute Mode (Accelerator): Cxl.cache or OpenCAPI C1



Memory Disaggregation

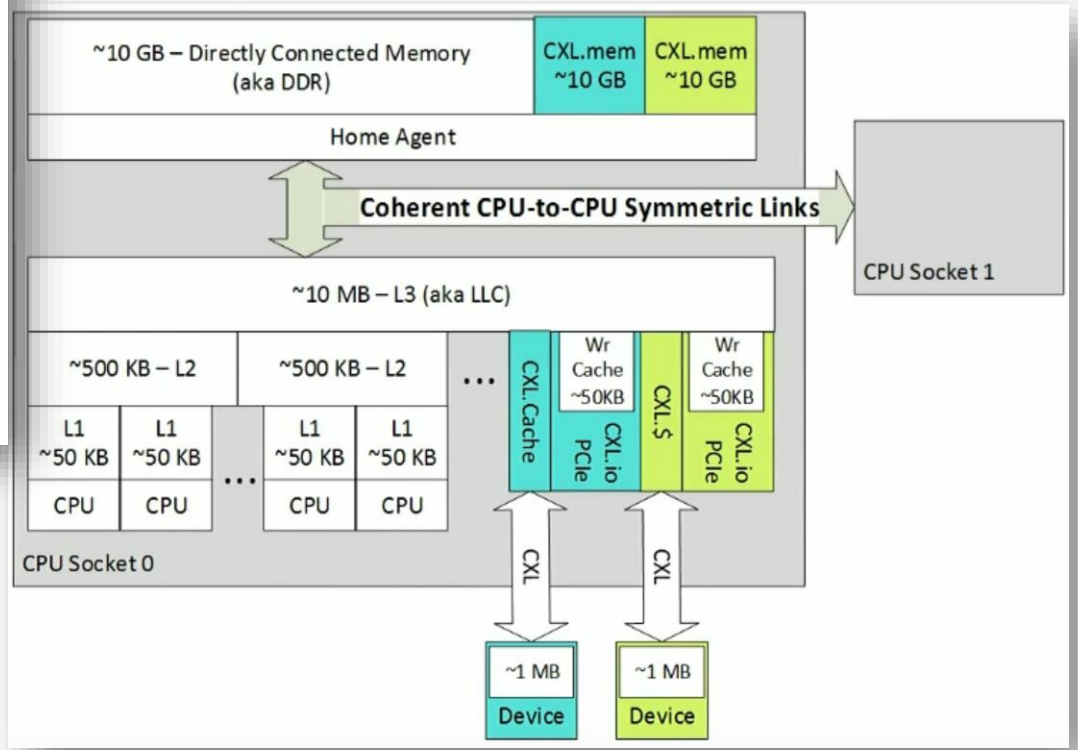
OSM Group

Difference OpenCAPI <-> CXL

2.3 TLX AP command packets

TLX commands are sent from the AFU to the host. An alphabetical list of the TLX commands follows; each command is hyperlinked to its specification. In this section, the TLX command specifications are in opcode order.

amo_rd	amo_rw	amo_w	assign_actag
dma_pr_w	dma_w	dma_w.be	intrp_req
intrp_req.d	nop	pr_rd_wnitc	rd_wnitc
wake_host_thread	xlate_touch		



Memory Disaggregation

OSM Group

- OpenCAPI
  - Read **w**ith **n**o **i**ntent **t**o **c**ache
  - Atomics (amo_*) supported
  - Accelerator only has non-coherent scratchpad
- CXL
  - Accelerator can have coherent cache
  - Integrated into a simpler MESI-based coherency domain

ThymesisFlow
Prototype in the SCORE-Lab

- 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

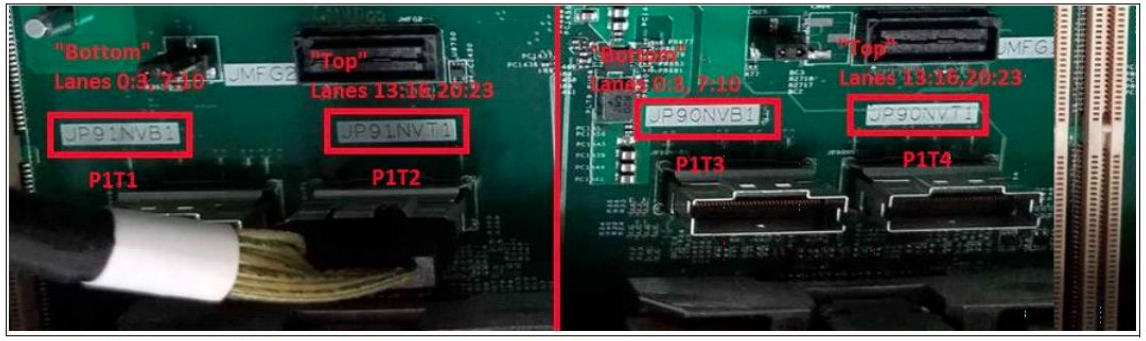
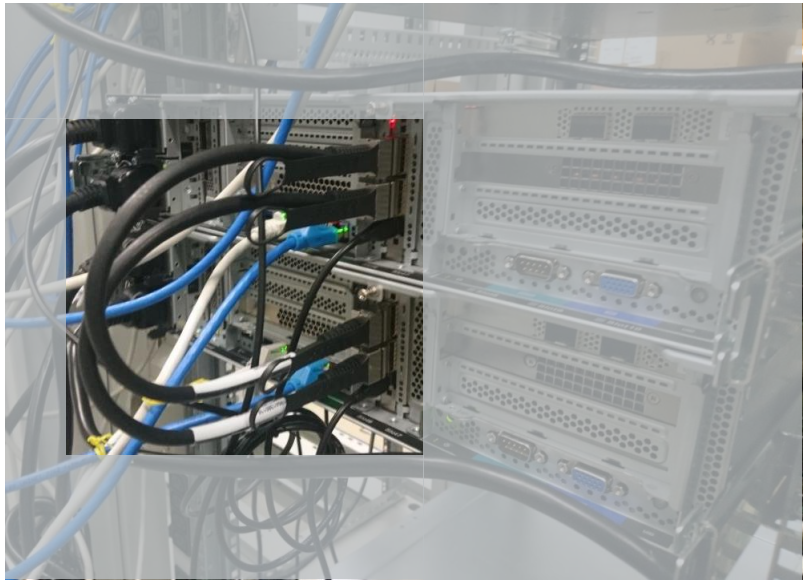


Figure 1-9 OpenCAPI 3.0 ports in the Power IC922 server

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
```

Memory Disaggregation

OSM Group

ThymesisFlow
Prototype in the SCORE-Lab

- ThymesisFlow (compute and memory mode) AFUs
  - https://github.com/OpenCAPI/ThymesisFlow
- LibtyhmesisFlow – Agent to configure FPGA AFUs
  - https://github.com/OpenCAPI/ThymesisFlow/tree/master/libthymesisflow
- Firmware patches for hotplugging (integration into Linux)
  - https://github.com/open-power/skiboot
  - Creates memory bars (with physical address = static offsets) in the device tree
- Kernel Module (mishmem) exposes static offsets as device to be mmaped
- Linux Kernel > 5.x
  - Then: RedHat 8.3 with 4.18.0-240.22.1.el8_3.ppc64le
  - Issues with SpectrumScale kernel module
  - Now: Ubuntu 20.04 with 5.4.0-81-generic

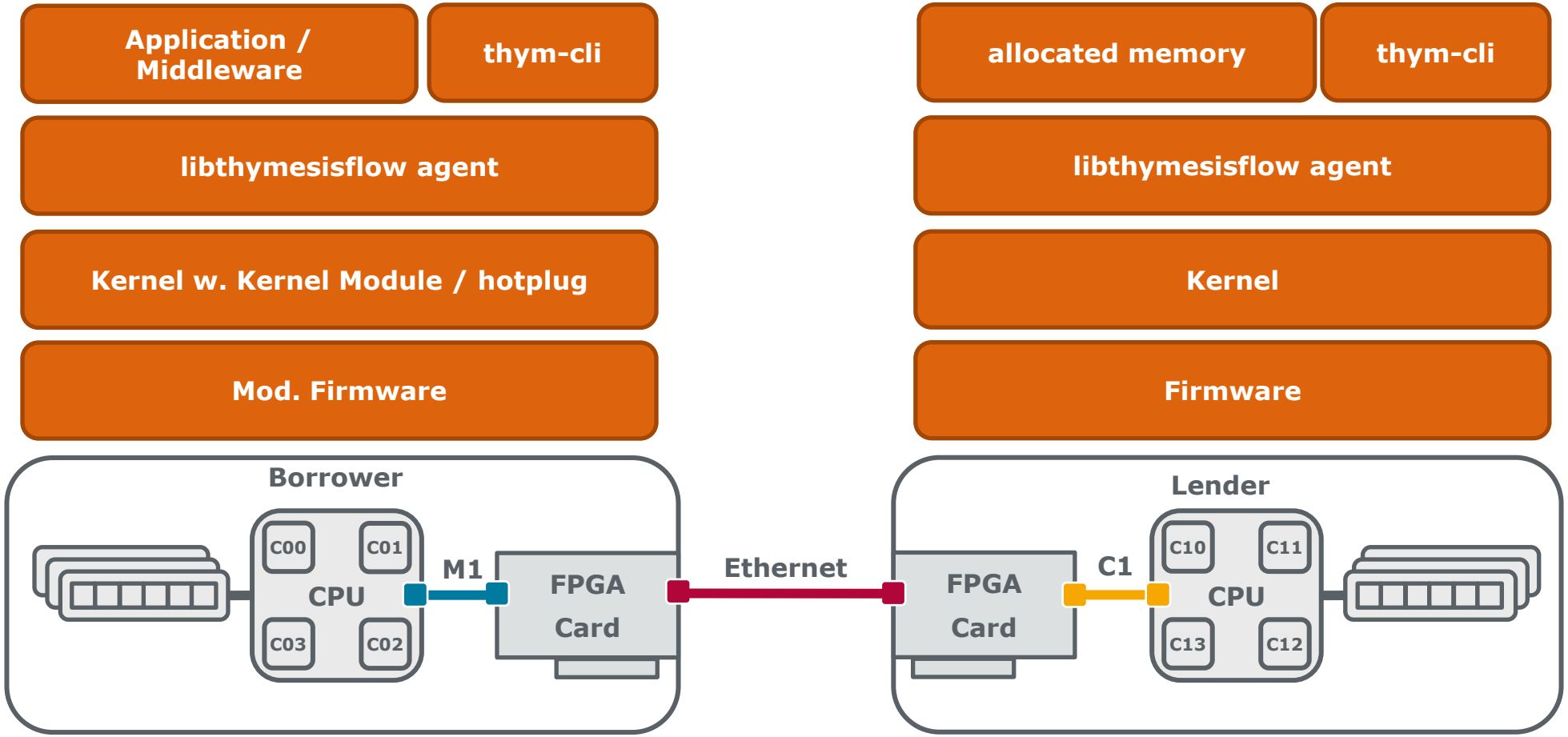
Setup Compute side

On the **Compute node** we can now attach the remote memory as following:

```
sudo ./bin/thymesisf-cli attach-compute \
    --afu <AFU_NAME> \
    --cid <CIRCUIT_ID> \
    --size <MEMORY_SIZE> \
    --ea <EFFECTIVE_ADDRESS> \
    [--no-hotplug]
```

On the **Memory node** you can reserve memory as following:

```
sudo ./bin/thymesisf-cli attach-memory \
    --afu <AFU_NAME> \
    --cid <CIRCUIT_ID> \
    --size <MEMORY_SIZE>
```



Memory Disaggregation

OSM Group

Latency Measurements

- Memory available either
  - /dev/mishmem mmap'able
  - NUMA-node if hotplugged
- Pointer-c
  - 128-by

```
static void *map_buf(off_t offset, uint64_t size) {
        int fd;
        void *ret;

        if ((fd = open("/dev/mishmem", O_RDWR|O_SYNC
                printf("\nError opening /dev/mishmem
                close(fd);
                return NULL;
        }
```

```
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:
```

```
[felix.eberhardt@ic922-01 libthymesisflow]$ bin/thymesisf-agent -s /tmp/thymesis.sock
2021-07-01 16:01:54 [INFO] using socket: /tmp/thymesisflow.sock
2021-07-01 16:01:54 [INFO] starting server with sock_path: /tmp/thymesisflow.sock
2021-07-01 16:01:54 [INFO] Starting thymesisflow server...
2021-07-01 16:01:54 [INFO] Ready to accept new requests...
2021-07-01 16:02:18 [INFO] allocating compute side...
```

```
[   22.914575] ocxl 0006:00:00.0: invalid short VPD tag 00 at offset 8452
[  528.900162] Starting MISHMEM device driver initialization
[  528.900887] "mishmem" character device successfully created
[  604.190492] MCE: CPU47: machine check (Severe) Host UE Load/Store DAR: 00007fffb0b80000 paddr: ffffffffffff0000 [Not recovered]
[  604.190493] MCE: CPU47: PID: 8921 Comm: test_compute_si NIP: [0000000010000a4c]
[  604.190493] MCE: CPU47: Initiator CPU
[  604.190493] MCE: CPU47: Hardware error
[  604.190494] test_compute_si[8921]: bus error (7) at 10000a4c nip 10000a4c lr 100009e0 code 4 in test_compute_single[10000000+10000]
[  604.190495] test_compute_si[8921]: code: 39290001 913f0060 813f0060 2f8901ff 409dffc4 39200000 913f0060 48000030
[  604.190496] test_compute_si[8921]: code: e93f0062 79291764 e95f0078 7d2a4a14 <81290000> 815f0068 7d2a4a14 913f0068
[  604.190560] Memory failure: 0xffffffffffff: memory outside kernel control
[  604.198551] EEH: dead PHB#6 detected, location: N/A
[  604.198568] EEH: Beginning: 'error_detected(permanent failure)'
```

Latency Measurements

Pointer-chasing workload

Elements	Total Size	Avg Latency / element
83.554.432	10.199 MB	854 ns
123.554.432	15.082 MB	867 ns
203.554.432	24.847 MB	913 ns
303.554.432	37.054 MB	947 ns
383.554.432	46.820 MB	975 ns

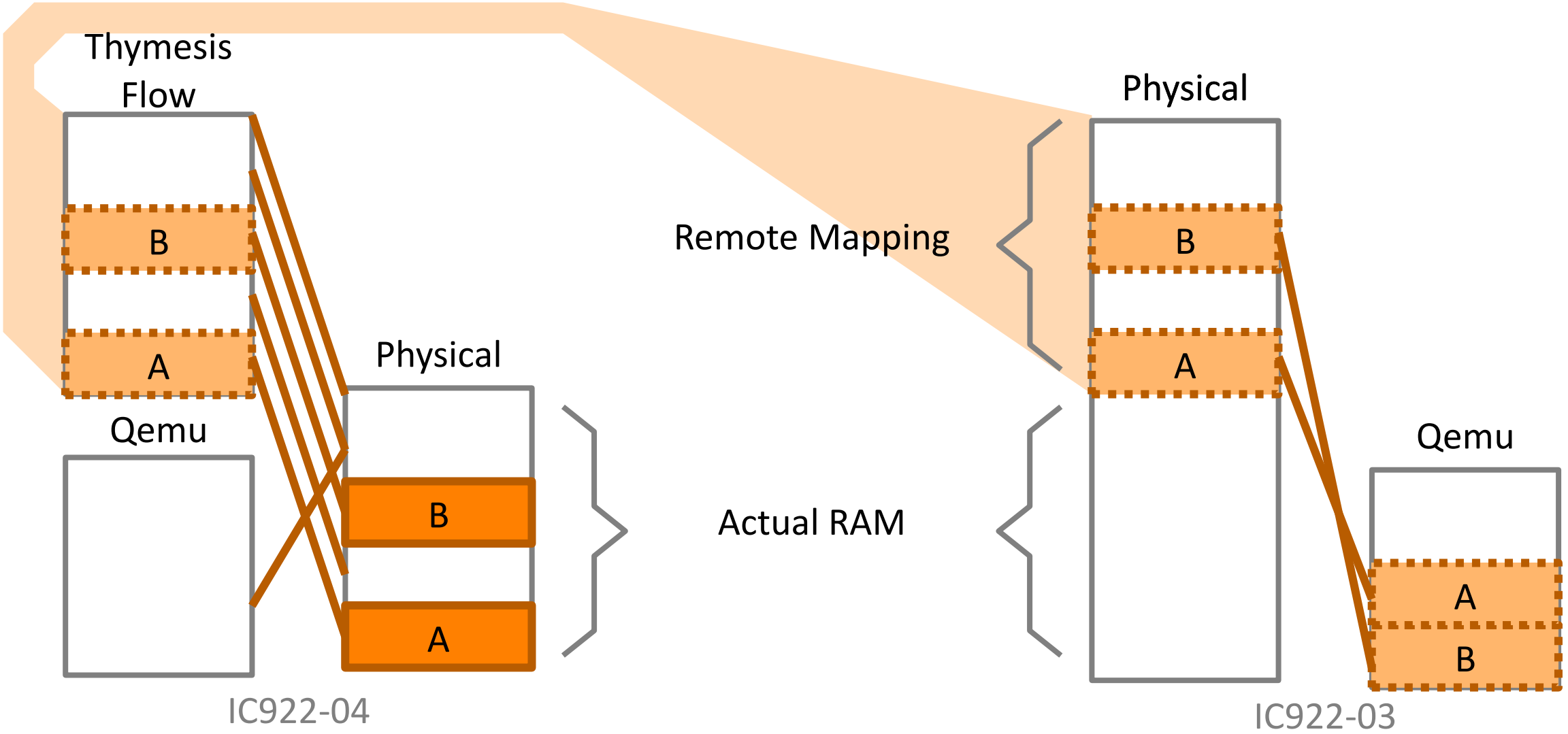
Power9 processor has a total of 120 MB L3 cache for all cores, Page size 64K

More Information at:

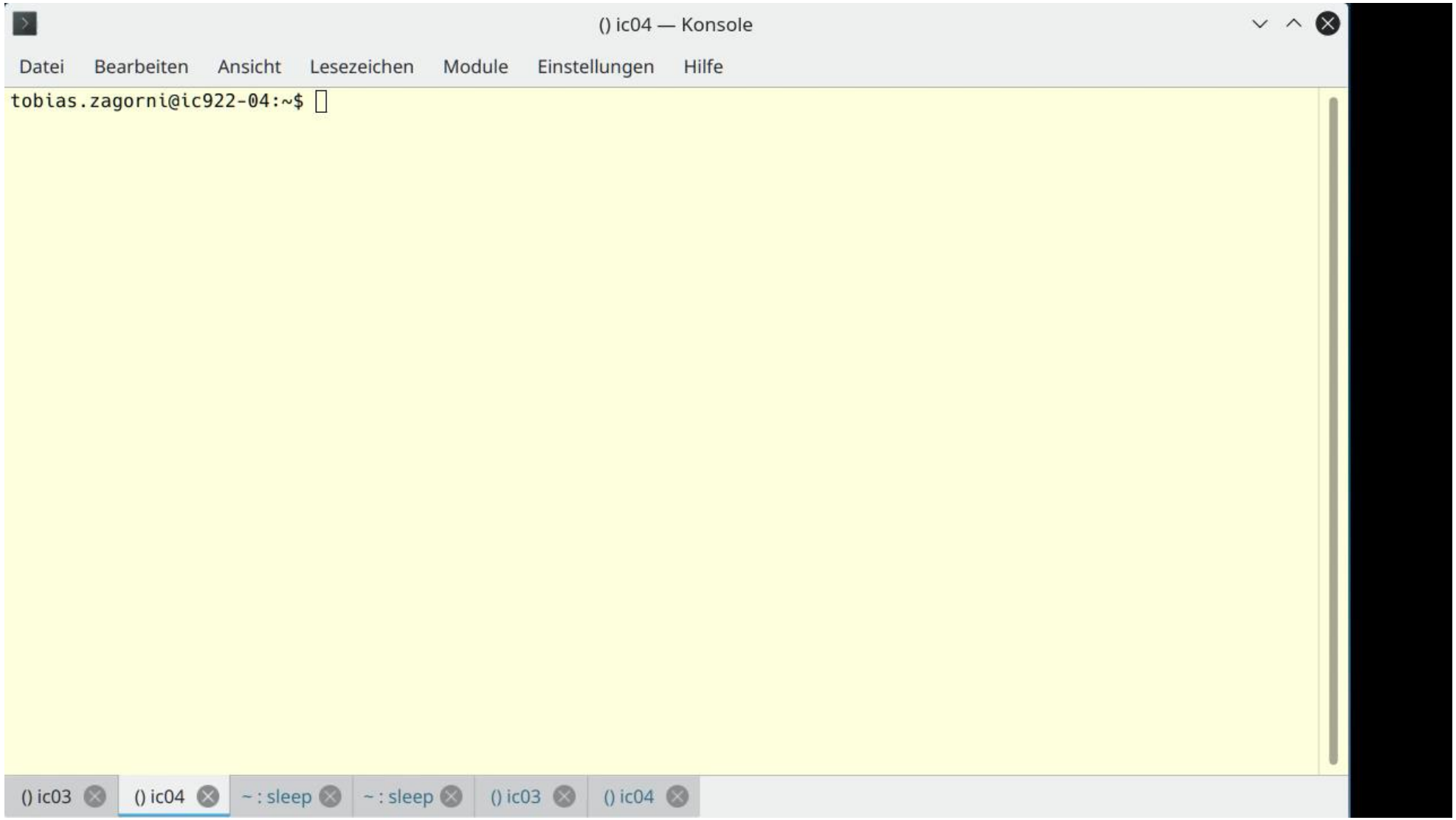
ThymesisFlow: A Software-Defined, HW/SW co-Designed Interconnect Stack for Rack-Scale Memory Disaggregation

https://www.microarch.org/micro53/papers/738300a868.pdf

Demo
Migrate VM from IC922-04 to IC922-03



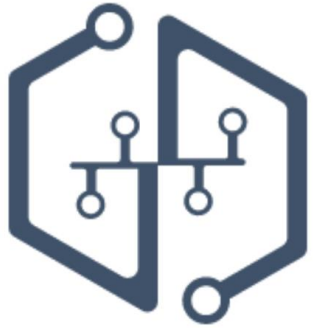
Demo
VM migration with QEMU



Invitation for Collaboration / Contribution Workshop / Lab Resources

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

COMPSYS '22



Workshop on Composable Systems
Co-located with IPDPS 2022

Important dates

- Papers submissions: February 1 2022
- Authors notification: March 1 2022
- Camera ready: March 15 2022
- Workshop: May 30 through June 3 2022

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