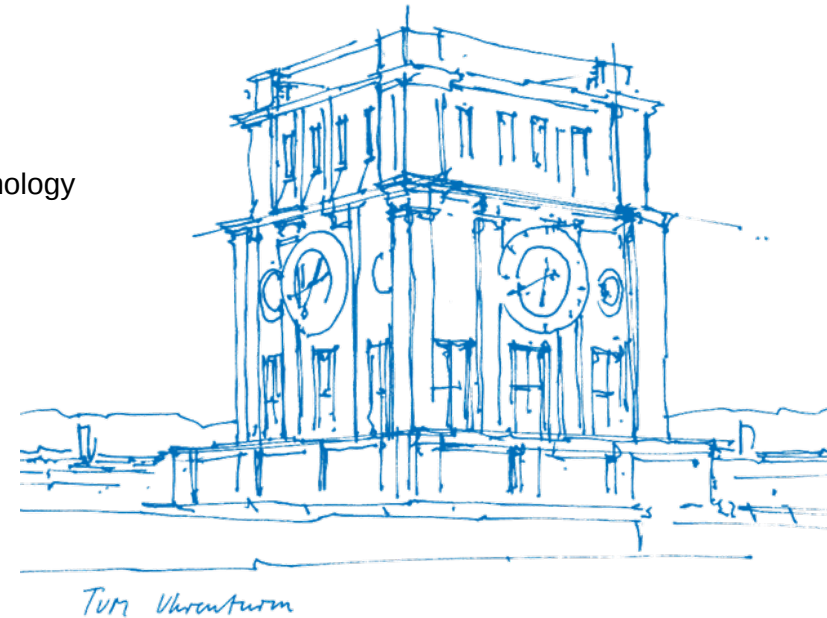


Chiplet Technology and the Impact of NUMA on Applications

Alex Daichendt

IN2076 Advanced Computer Architecture
School of Computation, Information, and Technology
Technical University of Munich



Intel® Data Center Advantages vs. AMD Naples

Implementation Matters

INTEL

HOLISTIC DATA CENTER SOLUTIONS

- ✓ Proven System Performance and Innovation
- ✓ Architected for the Data Center
- ✓ Robust SW and HW Ecosystem

AMD

REPURPOSED DESKTOP PRODUCT FOR SERVER

- ✗ Inconsistent Performance from 4 Glued-together Desktop Die
- ✗ Poor Track Record, Inconsistent Supplier
- ✗ Lack of Ecosystem

¹TechPowerUp; no primary source available

Why chiplets?



- Moore's Law
- more flexibility in design

²[1] S. Naffziger *et al.*, "Pioneering Chiplet Technology and Design for the AMD EPYC™ and Ryzen™ Processor Families : Industrial Product", IEEE, Jun. 2021. doi: 10.1109/isca52012.2021.00014.

Why chiplets?

- Moore's Law
- more flexibility in design
- low production yield for monolithic dies
→ \$\$\$

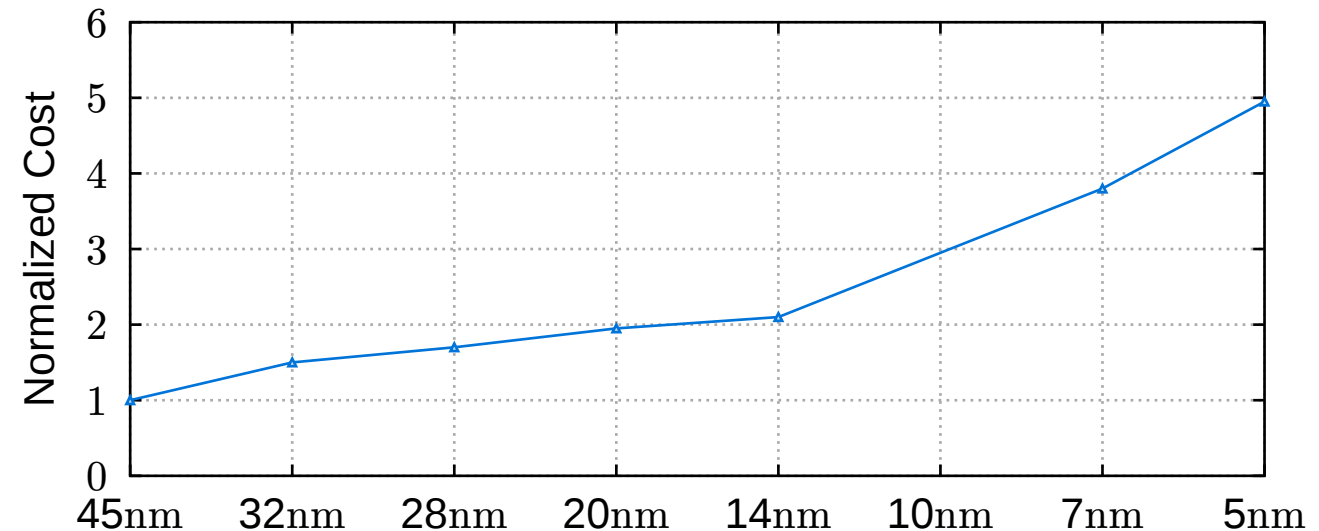
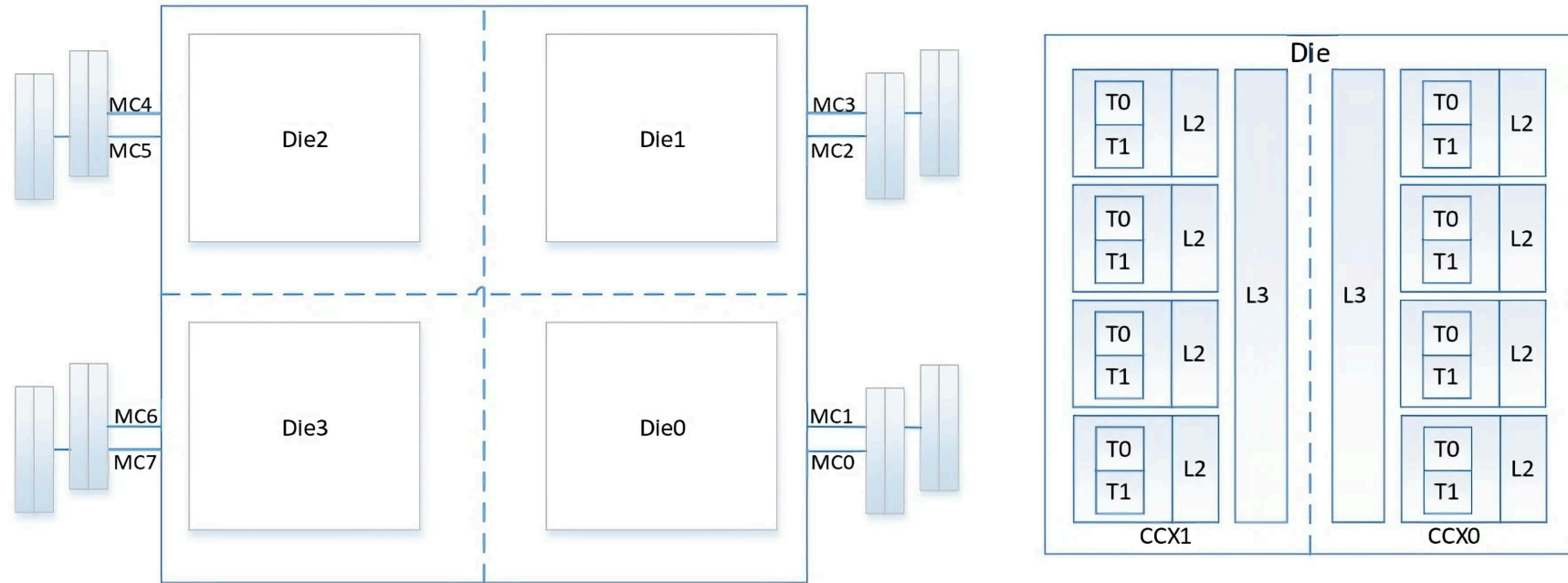


Figure 2: Normalized cost per chip vs. technology node, based on Naffziger et al.³

³[1] S. Naffziger et al., "Pioneering Chiplet Technology and Design for the AMD EPYC™ and Ryzen™ Processor Families : Industrial Product", IEEE, Jun. 2021. doi: 10.1109/isca52012.2021.00014.

AMD Naples (1st Gen. EPYC) – NUMA Topology⁴



Legend

Tx = thread

Lx = L2/3 cache

MCx = Memory Controller

CCXx = Core Complex

⁴<https://www.amd.com/content/dam/amd/en/documents/epyc-technical-docs/specifications/56308-numa-topology-for-epyc-naples-family-processors.pdf>

AMD Naples (1st Gen. EPYC)

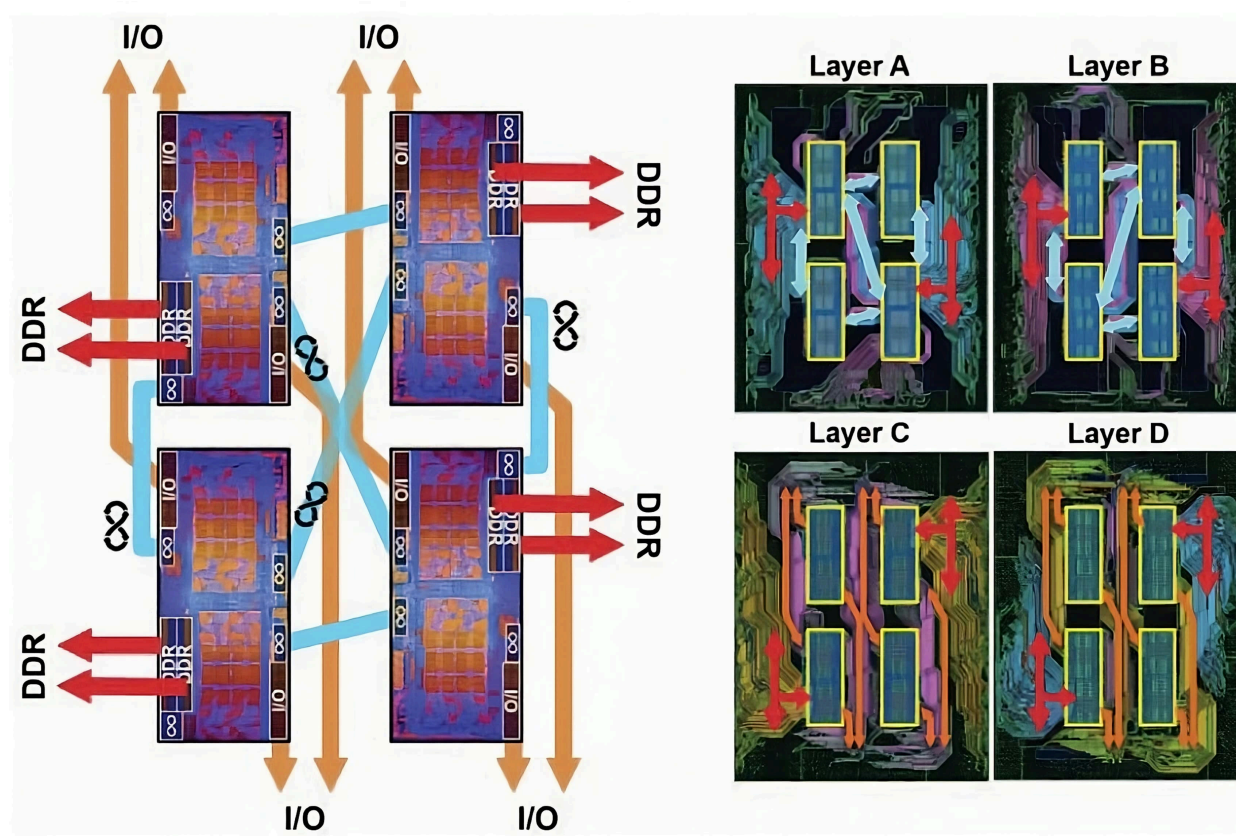
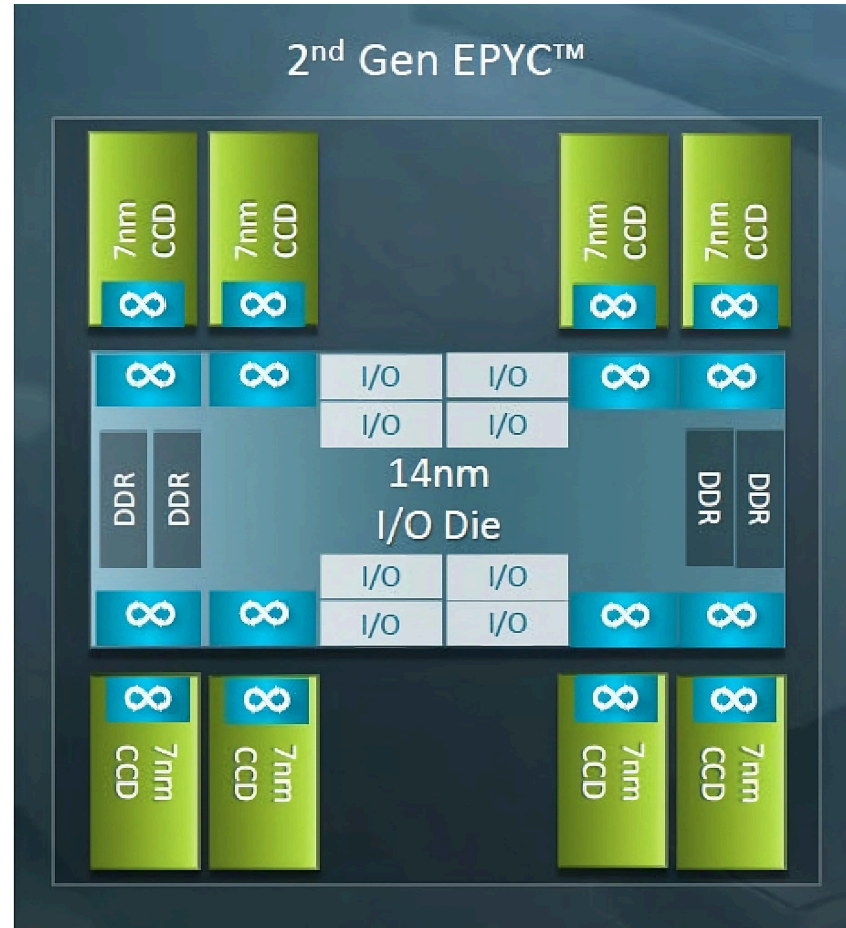


Figure 4: Multi-layer package routing, DDR (red), IO (orange), infinity-fabric (blue)⁵

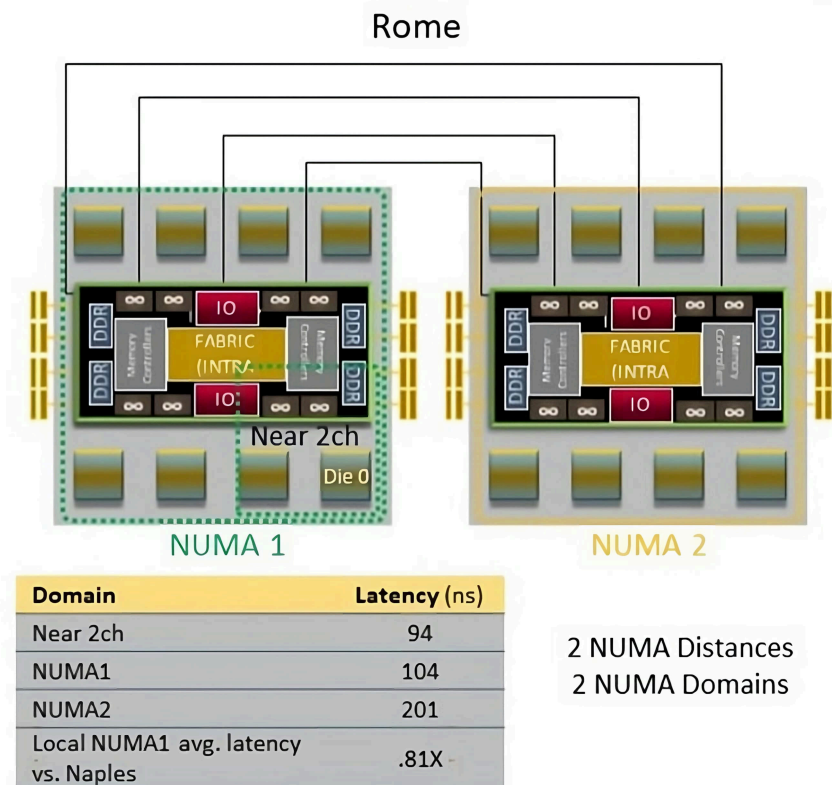
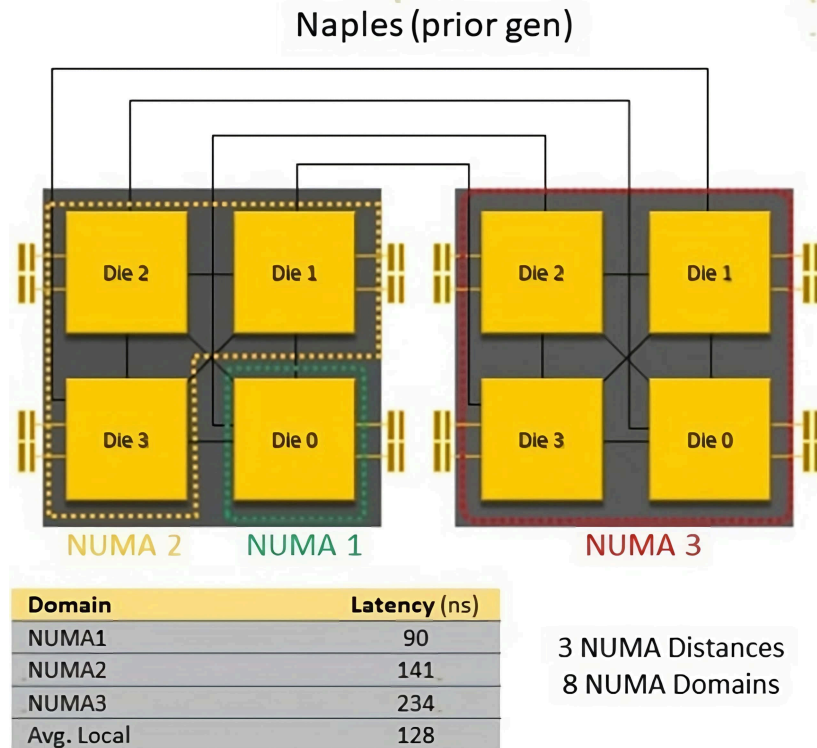
⁵[1] S. Naffziger *et al.*, "Pioneering Chiplet Technology and Design for the AMD EPYC™ and Ryzen™ Processor Families : Industrial Product", IEEE, Jun. 2021. doi: 10.1109/isca52012.2021.00014.

AMD Rome (2nd Gen. EPYC)⁶



⁶ <https://www.amd.com/content/dam/amd/en/documents/epyc-technical-docs/tuning-guides/amd-epyc-7002-tg-hpc-56827.pdf>

Memory Access Latencies for Naples and Rome, Naffziger et al.⁷



⁷ [2] S. Naffziger, K. Lepak, M. Paraschou, and M. Subramony, "2.2 AMD Chiplet Architecture for High-Performance Server and Desktop Products", IEEE, Feb. 2020. doi: 10.1109/isscc19947.2020.9063103.

Impact of NUMA on Applications

Emmerich et al.⁸ – User Space Networking Drivers

Ingress NIC	Egress NIC	CPU	Memory	Throughput
Node 0	Node 0	Node 0	Node 0	10.8M pps
Node 0	Node 0	Node 0	Node 1	10.8M pps
Node 0	Node 0	Node 1	Node 0	7.6M pps
Node 0	Node 0	Node 1	Node 1	6.6M pps
Node 0	Node 1	Node 0	Node 0	7.9M pps
Node 0	Node 1	Node 0	Node 1	10.0M pps
Node 0	Node 1	Node 1	Node 0	8.6M pps
Node 0	Node 1	Node 1	Node 1	8.1M pps

Figure 6: Forwarding performance, columns indicates pinning of resources, based on Emmerich et al.[3]

⁸[3] P. Emmerich, M. Pudelko, S. Bauer, and G. Carle, “User Space Network Drivers”, in ANRW '18. Montreal, QC, Canada: Association for Computing Machinery, 2018, p. 91. doi: 10.1145/3232755.3232767.

Impact of NUMA on Applications

Li et al.⁹ – Memcached

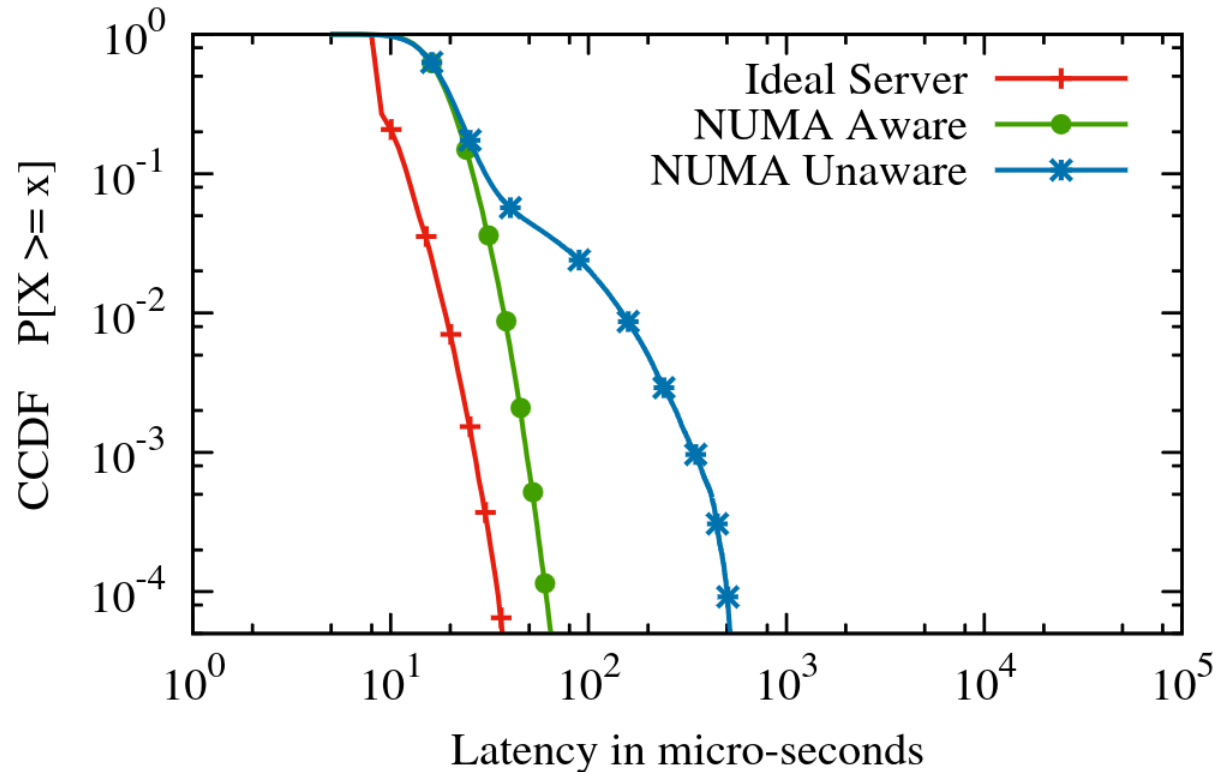


Figure 7: Memcached tail latency; 2 sockets; two instances (green), one instance (blue), based on Li et al.[4]

⁹[4] J. Li, N. K. Sharma, D. R. K. Ports, and S. D. Gribble, "Tales of the Tail", ACM, Nov. 2014. doi: 10.1145/2670979.2670988.

Conclusion



- CPU architecture matters
- Chiplet technology is a fundamental part of future CPU architectures
- Inconsistent memory access latencies are a challenge for applications

- [1] S. Naffziger *et al.*, “Pioneering Chiplet Technology and Design for the AMD EPYC™ and Ryzen™ Processor Families : Industrial Product”, IEEE, Jun. 2021. doi: 10.1109/isca52012.2021.00014.
- [2] S. Naffziger, K. Lepak, M. Paraschou, and M. Subramony, “2.2 AMD Chiplet Architecture for High-Performance Server and Desktop Products”, IEEE, Feb. 2020. doi: 10.1109/isscc19947.2020.9063103.
- [3] P. Emmerich, M. Pudelko, S. Bauer, and G. Carle, “User Space Network Drivers”, in ANRW '18. Montreal, QC, Canada: Association for Computing Machinery, 2018, p. 91. doi: 10.1145/3232755.3232767.
- [4] J. Li, N. K. Sharma, D. R. K. Ports, and S. D. Gribble, “Tales of the Tail”, ACM, Nov. 2014. doi: 10.1145/2670979.2670988.