Design and implementation of a distributed cache-coherent memory

— Advisors : Gaël Thomas
— Location : Benagil team, Telecom SudParis/Inria, Palaiseau building, and Whisper Team, Inria Building, Paris

1 Context

Resource disaggregation has recently attracted significant interest as a solution to simplify resource management in a data center [1, 2, 12, 13, 15, 17, 18, 21, 25, 28, 31, 35, 36, 38, 39]. It consists of designing a data center with physical machines specialized in a single type of resource (CPU nodes, memory nodes, GPU nodes...). These machines are interconnected via high-speed networks, which enables a cloud provider to create a virtual machine on demand by re-aggregating disaggregated resources.

In a disaggregated data center, remote memory access is slower than local memory access [11, 14, 32]. To hide the increased latency, both industry and academia propose to use local memory on the CPU node as a software-managed cache for remote memory [12, 12, 21, 28, 35, 39]. In this setting, hot data is kept on the CPU node, while cold data remains on the remote memory node.

Transparently executing an application on top of a disaggregated infrastructure remains a challenge. Current runtimes are implemented either inside the kernel as a swap device [12] or by relying on specific language features [28]. Neither of these solutions is satisfactory. The swap device approach is fully transparent but not flexible enough, since it is difficult to finely tune the Linux cache replacement policy to the specific needs of an application. The language-based approach is more flexible, as the replacement policy is directly embedded in the application. However, in this case, remote memory access is no longer transparent, because it requires the use of special data structures.

2 Subject

With VoliMem, we propose a runtime that implements disaggregation while achieving both transparency and flexibility. Instead of implementing the disaggregated runtime directly in the kernel, we propose to implement it in user space by relying on the virtualization instruction set of a modern processor. Thanks to virtualization, VoliMem creates a second page table within the process, which is used to transparently implement swapping in user space. Because VoliMem leverages a page table, it is fully transparent to the application. Moreover, since swapping is implemented in user space, the replacement policy can be finely tuned to the specific needs of the application.

In its current state, a process running in VoliMem can aggregate memory from several memory nodes. However, the process cannot yet run across multiple CPU nodes because VoliMem does not currently implement a cache-coherency protocol. The goal of this internship is therefore to explore how a cache-coherency protocol could be implemented in VoliMem, inspired by existing hardware protocols (i.e., MESIF and MOESI).

Work plan.

Month 1 : Related work,
Month 2-3 : Design and implementation of the cache-coherency protocol
Month 4 : Evaluations and report writing.

3 Advisors expertise

Gaël Thomas (Benagil team) is an expert in systems. He has been working on NUMA architectures [4, 9, 10, 34], privacy [30, 37], performance analysis [5, 22], persistent memory [6,16], concurrent programming [19,20,26,27], bug analysis [23,24,29], and language runtime designs [3, 7, 8, 33].

4 Expected skills

The candidate must have a good background in system programming and C/C++.

Références

[1] Emmanuel Amaro, Christopher Branner-Augmon, Zhihong Luo, Amy Ousterhout, Marcos K. Aguilera, Aurojit Panda, Sylvia Ratnasamy, and Scott Shenker. Can far memory improve job throughput ? In *Proceedings of the EuroSys European Conference on Computer Systems, EuroSys '20*, EuroSys '20, New York, NY, USA, 2020. ACM.

[2] Emmanuel Amaro, Christopher Branner-Augmon, Zhihong Luo, Amy Ousterhout, Marcos K. Aguilera, Aurojit Panda, Sylvia Ratnasamy, and Scott Shenker. Can far memory improve job throughput ? In *Proceedings of the EuroSys European Conference on Computer Systems, EuroSys '20*, EuroSys '20, New York, NY, USA, 2020. ACM.

[3] Koutheir Attouchi, Gaël Thomas, Gilles Muller, Julia Lawall, and André Bottaro. Incinerator - eliminating stale references in dynamic OSGi applications. In *Proceedings of the international conference on Dependable Systems and Networks, DSN'15*, page 11, Rio de Janeiro, Brazil, 2015. IEEE Computer Society.

[4] Bao Bui, Djob Mvondo, Boris Teabe, Kevin Jiokeng, Lavoisier Wapet, Alain Tchana, Gaël Thomas, Daniel Hagimont, Gilles Muller, and Noel De Palma. When eXtended para-virtualization (XPV) meets NUMA. In *Proceedings of the EuroSys European Conference on Computer Systems, EuroSys'19*, page 15, Dresden, Germany, 2019. ACM.

[5] Florian David, Gaël Thomas, Julia Lawall, and Gilles Muller. Continuously measuring critical section pressure with the Free-Lunch profiler. In *Proceedings of the conference on Object Oriented Programming Systems Languages and Applications, OOPSLA'14*, page 14, Portland, Oregon, US, 2014. ACM.

[6] Rémi Dulong, Rafael Pires, Andreia Correia, Valerio Schiavoni, Pedro Ramalhete, Pascal Felber, and Gaël Thomas. NVCache : A plug-and-play NVMM-based I/O booster for legacy systems. In *Proceedings of the international conference on Dependable Systems and Networks, DSN'21*, page 13, Taipei, Taiwan, 2021. IEEE Computer Society.

[7] Nicolas Geoffray, Gaël Thomas, Julia Lawall, Gilles Muller, and Bertil Folliot. VMKit : a substrate for managed runtime environments. In *Proceedings of the international conference on Virtual Execution Environments, VEE'10*, pages 51–62, Pittsburgh, PA, USA, 2010. ACM.

[8] Nicolas Geoffray, Gaël Thomas, Gilles Muller, Pierre Parrend, Stéphane Frénot, and Bertil Folliot. I-JVM : a java virtual machine for component isolation in OSGi. In *Proceedings of the international conference on Dependable Systems*

and Networks, DSN'09, pages 544–553, Estoril, Portugal, 2009. IEEE Computer Society.

[9] Lokesh Gidra, Gaël Thomas, Julien Sopena, and Marc Shapiro. A study of the scalability of stop-the-world garbage collectors on multicores. In *Proceedings of the conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS'13*, pages 229–240, Houston, Texas, USA, 2013. ACM.

[10] Lokesh Gidra, Gaël Thomas, Julien Sopena, Marc Shapiro, and Nhan Nguyen. NumaGiC : a garbage collector for big data on big NUMA machines. In *Proceedings of the conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS'15*, page 14, Istanbul, Turkey, 2015. ACM.

[11] Donghyun Gouk, Sangwon Lee, Miryeong Kwon, and Myoungsoo Jung. Direct access, High-Performance memory disaggregation with DirectCXL. In *Proceedings of the Usenix Annual Technical Conference, USENIX ATC '22*, pages 287–294, Carlsbad, CA, July 2022. USENIX Association.

[12] Juncheng Gu, Youngmoon Lee, Yiwen Zhang, Mosharaf Chowdhury, and Kang G. Shin. Efficient memory disaggregation with infiniswap. In *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*, pages 649–667, Boston, MA, March 2017. USENIX Association.

[13] Zhiyuan Guo, Yizhou Shan, Xuhao Luo, Yutong Huang, and Yiying Zhang. Clio : a hardware-software co-designed disaggregated memory system. In *Proceedings of the conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS '22*, ASPLOS '22, page 417–433, New York, NY, USA, 2022. ACM.

[14] Kishon Vijay Abraham I. PCI NTB Function, 2022. `https://www.kernel.org/doc/html/v6.1/PCI/endpoint/pci-ntb-function.html`.

[15] Andres Lagar-Cavilla, Junwhan Ahn, Suleiman Souhlal, Neha Agarwal, Radoslaw Burny, Shakeel Butt, Jichuan Chang, Ashwin Chaugule, Nan Deng, Junaid Shahid, Greg Thelen, Kamil Adam Yurtsever, Yu Zhao, and Parthasarathy Ranganathan. Software-defined far memory in warehouse-scale computers. In *Proceedings of the conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS '19*, ASPLOS '19, page 317–330, New York, NY, USA, 2019. ACM.

[16] Anatole Lefort, Yohan Pipereau, Kwabena Amponsem, Pierre Sutra, and Gaël Thomas. J-NVM : Off-heap persistent objects in java. In *Proceedings of the Symposium on Operating Systems Principles, SOSP'21*, page 16, online, 2021. ACM.

[17] Sergey Legtchenko, Nicholas Chen, Daniel Cletheroe, Antony Rowstron, Hugh Williams, and Xiaohan Zhao. Xfabric : a reconfigurable in-rack network for rack-scale computers. In *Proceedings of the conference on Networked Systems Design and Implementation, NSDI '16*, NSDI'16, page 15–29, USA, 2016. USENIX Association.

[18] Huaicheng Li, Daniel S. Berger, Lisa Hsu, Daniel Ernst, Pantea Zardoshti, Stanko Novakovic, Monish Shah, Samir Rajadnya, Scott Lee, Ishwar Agarwal, Mark D. Hill, Marcus Fontoura, and Ricardo Bianchini. Pond : Cxl-based memory pooling systems for cloud platforms. In *Proceedings of the conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS '23*, ASPLOS 2023, page 574–587, New York, NY, USA, 2023. ACM.

[19] Jean-Pierre Lozi, Florian David, Gaël Thomas, Julia Lawall, and Gilles Muller. Remote core locking : migrating critical-section execution to improve the performance of multithreaded applications. In *Proceedings of the Usenix Annual Technical Conference, USENIX ATC'12*, pages 65–76, Boston, MA, USA, 2012. USENIX Association.

[20] Jean-Pierre Lozi, Florian David, Gaël Thomas, Julia Lawall, and Gilles Muller. Fast and portable locking for multicore architectures. *ACM Transactions on Computer Systems (TOCS)*, 33(4) :13 :1–13 :62, January 2016.

[21] Haoran Ma, Shi Liu, Chenxi Wang, Yifan Qiao, Michael D. Bond, Stephen M. Blackburn, Miryung Kim, and Guoqing Harry Xu. Mako : a low-pause, high-throughput evacuating collector for memory-disaggregated datacenters. In *Proceedings of the conference on Programming Language Design and Implementation, PLDI '22*, PLDI 2022, page 92–107, New York, NY, USA, 2022. ACM.

[22] Mohamed Said Mosli, François Trahay, Alexis Lescouet, Gauthier Voron, Rémi Dulong, Amina Guermouche, Élisabeth Brunet, and Gaël Thomas. Using differential execution analysis to identify thread interference. *IEEE Transactions on Parallel and Distributed Systems (TPDS)*, 30(12) :13, 2019.

[23] Nicolas Palix, Gaël Thomas, Suman Saha, Christophe Calvès, Julia Lawall, and Gilles Muller. Faults in linux : ten years later. In *Proceedings of the conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS'11*, pages 305–318, Newport Beach, CA, USA, 2011. ACM.

[24] Nicolas Palix, Gaël Thomas, Suman Saha, Christophe Calvès, Gilles Muller, and Julia Lawall. Faults in linux 2.6. *ACM Transactions on Computer Systems (TOCS)*, 32(2) :4 :1–4 :40, 2014.

[25] Christian Pinto, Dimitris Syrivelis, Michele Gazzetti, Panos Koutsovasilis, Andrea Reale, Kostas Katrinis, and H. Peter Hofstee. Thymesisflow : A software-

defined, hw/sw co-designed interconnect stack for rack-scale memory disaggregation. In *Proceedings of the International Symposium on Microarchitecture, MICRO '20*, pages 868–880, 2020.

[26] Thomas Preud'Homme, Julien Sopena, Gaël Thomas, and Bertil Folliot. BatchQueue : fast and memory-thrifty core to core communication. In *Proceedings of the international Symposium on Computer Architecture and High Performance Computing, SBAC-PAD'10*, pages 215–222, Petrópolis, Brazil, 2010. IEEE Computer Society.

[27] Thomas Preud'homme, Julien Sopena, Gaël Thomas, and Bertil Folliot. An improvement of OpenMP pipeline parallelism with the BatchQueue algorithm. In *Proceedings of the International Conference on Parallel and Distributed Systems, ICPADS'12*, page 8, Singapore, 2012. IEEE Computer Society.

[28] Zhenyuan Ruan, Malte Schwarzkopf, Marcos K. Aguilera, and Adam Belay. AIFM : High-Performance, Application-Integrated far memory. In *Proceedings of the conference on Operating Systems Design and Implementation, OSDI '20*, pages 315–332. USENIX Association, November 2020.

[29] Suman Saha, Jean-Pierre Lozi, Gaël Thomas, Julia Lawall, and Gilles Muller. Hector : Detecting resource-release omission faults in error-handling code for systems software. In *Proceedings of the international conference on Dependable Systems and Networks, DSN'13*, page 12, Budapest, Hungary, 2013. IEEE Computer Society. **Best paper award**.

[30] Vasily A. Sartakov, Stefan Brenner, Sonia Ben Mokhtar, Sara Bouchenak, Gaël Thomas, and Rüdiger Kapitza. Eactors : Fast and flexible trusted computing using sgx. In *Proceedings of the International Conference on Middleware, Middleware'18*, page 12, Rennes, France, 2018. ACM.

[31] Yizhou Shan, Yutong Huang, Yilun Chen, and Yiying Zhang. LegoOS : a disseminated, distributed OS for hardware resource disaggregation. In *Proceedings of the conference on Operating Systems Design and Implementation, OSDI '18*, OSDI'18, page 69–87, USA, 2018. USENIX Association.

[32] Debendra Das Sharma. Compute Express Link (CXL) : Enabling heterogeneous data-centric computing with heterogeneous memory hierarchy. *IEEE Micro*, 2022.

[33] Gaël Thomas, Nicolas Geoffray, Charles Clément, and Bertil Folliot. Designing highly flexible virtual machines : the JnJVM experience. *Software - Practice & Experience (SP&E)*, 38(15) :1643–1675, 2008.

[34] Gauthier Voron, Gaël Thomas, Vivien Quéma, and Pierre Sens. An interface to implement NUMA policies in the xen hypervisor. In *Proceedings of the EuroSys*

European Conference on Computer Systems, EuroSys'17, page 14, Belgrade, Serbia, 2017. ACM.

[35] Chenxi Wang, Haoran Ma, Shi Liu, Yuanqi Li, Zhenyuan Ruan, Khanh Nguyen, Michael D. Bond, Ravi Netravali, Miryung Kim, and Guoqing Harry Xu. *Semeru : A Memory-Disaggregated Managed Runtime*. USENIX Association, USA, 2020.

[36] Xinan Yan, Bernard Wong, and Sharon Choy. R3s : Rdma-based rdd remote storage for spark. In *Proceedings of the 15th International Workshop on Adaptive and Reflective Middleware*, ARM 2016, New York, NY, USA, 2016. ACM.

[37] Peterson Yuhala, Jämes Ménétrey, Pascal Felber, Valerio Schiavoni, Alain Tchana, Gaël Thomas, Hugo Guiroux, and Jean-Pierre Lozi. Montsalvat : Intel SGX shielding for GraalVM native images. In *Proceedings of the International Conference on Middleware, Middleware'21*, page 13, Québec, Canada, 2021. ACM.

[38] Jin Zhang, Zhuocheng Ding, Yubin Chen, Xingguo Jia, Boshi Yu, Zhengwei Qi, and Haibing Guan. Giantvm : A type-ii hypervisor implementing many-to-one virtualization. In *Proceedings of the international conference on Virtual Execution Environments, VEE '20*, VEE '20, page 30–44, New York, NY, USA, 2020. ACM.

[39] Yang Zhou, Hassan M. G. Wassel, Sihang Liu, Jiaqi Gao, James Mickens, Minlan Yu, Chris Kennelly, Paul Turner, David E. Culler, Henry M. Levy, and Amin Vahdat. Carbink : Fault-Tolerant far memory. In *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*, pages 55–71, Carlsbad, CA, July 2022. USENIX Association.