# Monitoring Memory Behaviors and Mitigating NUMA Drawbacks on Tiered NVM Systems

Shengjie Yang, Xinyu Li, Xinglei Dou, Xiaoli Gong, Hao Liu, Li Chen, Lei Liu

▶ **To cite this version:**

Shengjie Yang, Xinyu Li, Xinglei Dou, Xiaoli Gong, Hao Liu, et al.. Monitoring Memory Behaviors and Mitigating NUMA Drawbacks on Tiered NVM Systems. 17th IFIP International Conference on Network and Parallel Computing (NPC), Sep 2020, Zhengzhou, China. pp.386-391, 10.1007/978-3-030-79478-1_33 . hal-03768756

**HAL Id: hal-03768756**
**https://inria.hal.science/hal-03768756**

Submitted on 4 Sep 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License


This document is the original author manuscript of a paper submitted to an IFIP conference proceedings or other IFIP publication by Springer Nature. As such, there may be some differences in the official published version of the paper. Such differences, if any, are usually due to reformatting during preparation for publication or minor corrections made by the author(s) during final proofreading of the publication manuscript.

# Monitoring Memory Behaviors and Mitigating NUMA Drawbacks on Tiered NVM Systems

Shengjie Yang[1,2], Xinyu Li[1,2], Xinglei Dou[1,2], Xiaoli Gong[3], Hao Liu[4], Li Chen[2], Lei Liu*[1,2]

[1]Sys-Inventor Lab, [2]SKLCA, ICT, CAS; [3]Nankai; [4]AMS, PLA & Tsinghua

**Abstract.** Non-Volatile Memory with byte-addressability invites a new paradigm to access persistent data directly. However, this paradigm brings new challenges to the Non-Uniform Memory Access (NUMA) architecture. Since data accesses cross NUMA node can incur significant performance loss, and, traditionally, OS moves data to the NUMA node where the process accessing it locates to reduce the access latency. However, we find challenges when migrating data on NVM, which motivates us to migrate the process instead. We propose SysMon-N, an OS-level sampling module, to obtain access information about NVM in low overhead. Furthermore, we propose N-Policy to utilize the data collected by SysMon-N to guide process migration. We evaluate SysMon-N and N-Policy on off-the-shelf NVM devices. The experimental results show that they provide 5.9% to 3.62x bandwidth improvement in the case where cross-node memory accesses happen.

**Keywords:** DRAM-NVM, NUMA, OS, Migration, Scheduling

## 1 Introduction

Non-Volatile Memory (NVM) attaches to the memory bus promises DRAM-like latency, byte-addressability, and data persistence. NVM will become commonplace soon. Previous studies (e.g., [4]), focusing on kernel-bypassing, redesign the file system dedicated to NVM for reducing software overheads stemmed from kernel involvement. A critical feature of this type of file system is the "direct access" (i.e., DAX) style interface through the *mmap*() system call, through which the user process can map the NVM-based file into its address space and access the file content directly by load/store instructions from user space [1]. Different from the on-demand paging data access, NVM possesses both byte-addressability and persistency, which allows user processes to access the persistent data directly. However, NVM is usually mounted on a specific node and forms a tiered/hybrid memory system with DRAM on NUMA servers, leading to the risk of cross-node accesses (i.e., remote access). Remote access may cause dramatic performance degradation, and there are many studies to provide shreds of evidence for this.

In terms of the performance loss due to the remote accessing on DRAM, previous work moves data from remote NUMA node to the local node where the user process is running on. However, we find some challenges in the previous

This project is supported by the National Key Research and Development Program of China under Grant No.2017YFB1001602 and the NSFC under grants No.61502452, 61902206, 61702286. This work originates from L. Liu's series of studies in ISCA,PACT,TPDS,TC,etc.[5-12] on memory systems conducted in Sys-Inventor Lab. More details refer to Sys-Inventor Lab - https://liulei-sys-inventor.github.io. *Corresponding author (PI):lei.liu@zoho.com;liulei2010@ict.ac.cn

studies about the NVM-based systems. (1) There is no "struct page" for persistent data in NVM that managed by the DAX-aware file systems [1], leading to the complexity of page migration on the system using both DRAM and NVM. (2) Since the data blocks to be migrated are persistent, the process of page migration needs to be guaranteed as atomic and consistent using a transaction-like mechanism, which will introduce extra overheads on the critical path. (3) The persistent data usually has a much larger size than the volatile data, and frequent migrating of them will produce significant overheads [13]. These challenges motivate us to seek a new design.

In this work, we propose an new mechanism. Instead of moving persistent data, we migrate the process to the original node where the persistent data locates. In order to achieve our goal, we propose SysMon-N and N-Policy. SysMon-N is an OS-level memory behavior sampling module that can obtain the NVM access "hotness" (i.e., access times within a sampling interval) and the access mode (i.e., remote or local) for a user process with low overheads. N-Policy is a process migration policy designed for the user processes which use MVM. For instance, N-Policy reduces the expensive remote accesses to NVM by migrating the process to the node that is close to NVM. The experimental results show that SysMon-N and N-Policy can increase the bandwidth of read-intensive applications by 5.9% and the bandwidth of write-intensive applications by 2.71x to 3.62x when the incorrect core is allocated and remote access occurs.

## 2 The Art of Our Design

### 2.1 SysMon-N - Sampling Memory Systems with NVM

To tackle the problems mentioned above, we first design a practical OS-level memory behavior sampling module to capture the NVM access information. Our prior efforts [5,6,8] propose SysMon as an OS-level memory behavior monitoring module. SysMon periodically checks the access bits in Page Table Entries (PTEs) to obtain the page hotness. However, merely checking PTEs can not distinguish whether the page is located in NVM or DRAM. So, we design SysMon-N, based on SysMon [5], to provide the physical address information of the data, it achieves two objectives. (1) Sampling pages in NVM to collect the page hotness information while avoiding sampling pages in DRAM to narrow down the sampling space; (2) Checking whether remote access occurs and collecting related data access information.

As a preprocessing step, SysMon-N collects the NUMA topology information of the platform by scanning ACPI static resource affinity table (SRAT), where the topology information of all processors and memories are stored. By checking the `ACPI_SRAT_MEM_NON_VOLATILE` flag of the SRAT entries, SysMon-N can get the range of physical address of all NVM devices. Usually, the physical locations of all pages in a Virtual Memory Aera (VMA) are the same. For a specific VMA, SysMon-N gets the physical address of VMA's start page and checks whether it falls in the physical address range of an NVM device. If so, it means that all pages of the VMA are on one specific NVM device, and it is necessary to traverse the VMA's memory address. Otherwise, the VMA is not in NVM and can be

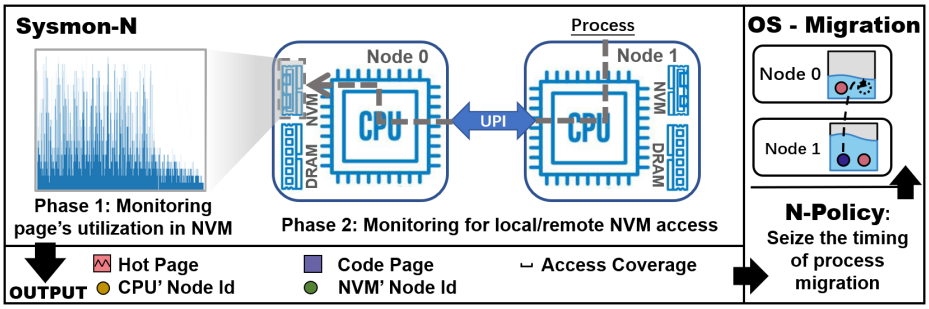



Fig. 1: Workflow of SysMon-N

skipped to narrow down the sampling space.

Figure 1 shows the workflow of SysMon-N. It has two phases. In phase 1, SysMon-N checks each page's `access_bit` within the monitored process to find the hot pages in NVM and their corresponding physical address area. Besides, considering the massive pressure that NVM's large capacity puts on the limited number of TLBs, it is natural for OS to use the huge page on NVM. The detection of the huge page utilization on NVM are basically the same with the 4KiB-based pages; SysMon-N uses the PMD entries to complete the address translation since the OS omits the last level PTE for 2MiB huge pages.

In phase 2, by comparing the *node id* of CPU where the process running on and that of the NVM node where the data are stored, SysMon-N can determine whether remote accesses occur or not. SysMon-N obtains the set of CPUs on which it's eligible to run by checking process's CPU affinity mask, and then calls the *cpu_to_node*() kernel function to check the node corresponding to the CPU. Finally, SysMon-N compared the CPU node id with NVM node id for the result: if the two node ids are the same, the process has accessed the page on remote NVM; otherwise, the process only touches the local NVM.

Finally, after sampling, SysMon-N has the number of hot and cold pages and related physical address ranges, and provides the information to N-Policy for making a decision.

### 2.2 N-Policy - For NVM

N-Policy leverages the formation provided by SysMon-N, and guides process migration accordingly. The key component of N-Policy is a *conditional migration*

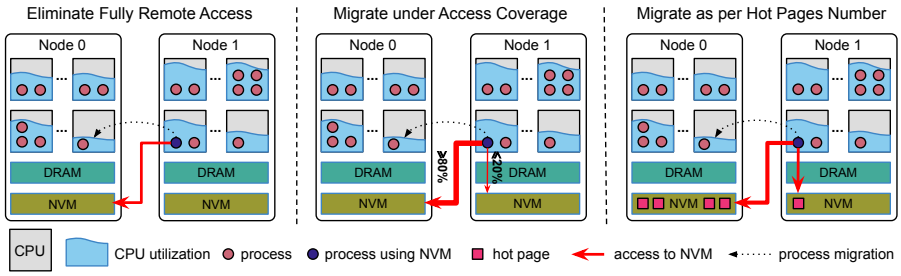



Fig. 2: Conditional Migration Model of N-Policy

*model* which is depicted in figure 2. It has two principles. (1) Eliminating remote access whenever possible; (2) Trying to avoid unnecessary migration. The inputs of N-Policy include: (1) the number of hot/cold pages on per node; (2) access coverage for each node; (3) CPU's node id on which the process locates; (4) node id for the used NVM.

After each SysMon-N sampling epoch is completed, the N-Policy immediately decides whether to migrate according to three data access conditions as shown in figure 2. The first case is when completely remote access occurs (no data are accessed from local NUMA node), the migration action is triggered to migrate the process to the same node of data. This situation is determined by judging whether there is intersection between the set of the node id of CPU which the process are allowed to run on and that of NVM where accessed data locates. N-Policy makes process migration in this case, easing the overhead of cross-node access by placing the process on the same node as the NVM being used.

If the two set of node ids have intersections, N-Policy compare *access coverage* on different nodes to further decide whether to execute migration. Access coverage symbolizes the amount of data accessed by the process on each NUMA node. If the access coverage of different nodes is unbalanced (i.e., in Figure 2, N-Policy considers access coverage on remote NUMA node greater than 80% as unbalanced access). N-Policy will select the least utilized CPU on the NUMA node with the broadest access coverage as the target of process migration.

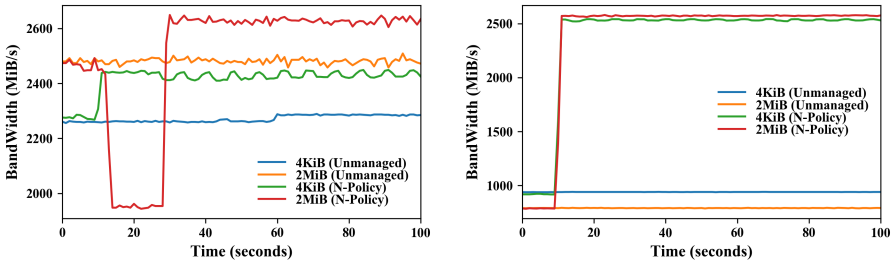
Finally, information about page hotness is also taken into account in N-Policy. Hot pages indicate frequently accessed pages and the data on them is often more important than other pages (may not be right in some cases), and should be accessed closer for reducing latency. To ensure fast access to hot pages, N-Policy compares NUMA node's hotness and migrate the process to the node with the more hot pages.

To avoid significant overheads caused by repeated and meaningless migrations, we let N-Policy receives messages from SysMon-N for every 10 seconds. N-Policy uses the function *sched_getaffinity*() of the Linux kernel to bind a process to the corresponding CPU nodes for migration.

## 3 Effectiveness of N-Policy on Bandwidth and Latency

Our experimental platform is a server with dual CPU sockets of Intel Xeon Gold 6240M CPU (each has 36 cores); it has 512GB Intel® Optane™ DC persistent memory on per socket, i.e., 1024GB NVM on our platform. We configure the namespace [3] for the Optane PMM, which represents a certain amount of NVM that can be formatted as logical blocks, and then deploy the ext4-DAX file system on it to support direct data access. We don't consider I/O in experiments [14].

We use the Flexible I/O Tester (Fio) [2] with `libpmem` engine to evaluate the effectiveness of N-policy collaborated SysMon-N. We adjust the minimum read/write block size of I/O operations to perform reads and writes to NVM in different situations, and record bandwidth under different block sizes with and without N-Policy enabled, respectively. To verify the effectiveness of N-Policy, all data accesses of Fio are set as remote access.




(a) 100% read (b) 100% write

Fig. 3: Unmanaged vs. Use N-Policy to guide migration

Figure 3-(a) presents the 100% read case. As a baseline, we launch Fio [2] in two cases with a single data access size of 4KiB and 2MiB, respectively. The corresponding memory bandwidth of the two cases is stable with an average of 2273 MiB/s and 2482 MiB/s, respectively. When N-Policy is enabled, it conducts process migration to eliminate the remote access which occurs at the timing around 10s. The bandwidth changes accordingly as the process is migrated to the optimal node. N-Policy can improve the bandwidth by 6.94% and 5.90% for 4KiB and 2MiB block size, respectively. Figure 3-(b) shows the 100% write case. N-Policy achieves better results in this case. By eliminating the remote access with process migration, the bandwidth of Fio can increase by 2.71x and 3.26x in the case of 4KiB and 2MiB block sizes, respectively. This is because reading and writing bandwidth on NVM are not symmetric and NVM is more sensitive to write operations.

## References

1. "Direct Access for files", `https://www.kernel.org/doc/Documentation/filesystems/dax.txt`.
2. "Fio - Flexible I/O tester", `https://fio.readthedocs.io/en/latest/`.
3. "Persistent Memory Concepts", `https://docs.pmem.io/ndctl-user-guide/concepts`.
4. D.S.Rao, et al, "System software for persistent memory", in EuroSys, 2014.
5. M.Xie, et al, "Sysmon: Monitoring memory behaviorsvia OS approach", in APPT, 2017.
6. L.Liu, et al, "Hierarchical Hybrid Memory Management in OS for Tiered Memory Systems", in IEEE TPDS, 2019.
7. X.Li, et al, "Thinking about A New Mechanism for Huge Page Management", in APSys, 2019.
8. L.Liu, et al, "Going Vertical in Memory Management: Handling Multiplicity by Multi-policy", in ISCA, 2014. (revised version)
9. L.Liu, et al, "BPM/BPM+: Software-based Dynamic Memory Partitioning Mechanisms for Mitigating DRAM Bank-/channel-level Interferences in Multicore Systems", in ACM TACO, 2014. (revised version)
10. L.Liu, et al, "A Software Memory Partition Approach for Eliminating Bank-level Interference in Multicore Systems", in PACT, 2012. (revised version)
11. L.Liu, et al, "Rethinking Memory Management in Modern Operating System: Horizontal, Vertical or Random?", in IEEE Trans. Computers (TC), 2016.
12. L.Liu, et al, "Memos: A Full Hierarchy Hybrid Memory Management Framework", in ICCD, 2016.
13. S.Chen, et al, "Efficient GPU NVRAM persistence with helper warps", in DAC, 2019.
14. F.Lv, et al, "Dynamic I/O-Aware Scheduling for Batch-Mode Applications on Chip Multiprocessor Systems of Cluster Platforms", in JCST, 2014.