Tiered Memory Systems with Virtual Persistent Memory

Dustin Nguyen
Wolfgang Schröder-Preikschat
FAU Erlangen-Nürnberg
Germany

Oliver Giersch
Jörg Nolte
BTU Cottbus-Senftenberg
Germany

Data centres make up a significant portion of the energy consumed globally. Just alone in the United States and Europe, they are responsible for around 4 % of the energy consumed, with an estimated growth of up to 6 % by 2026 [1]. In order to reduce the environmental impact of data centres, such as carbon emission, it is advisable to operate them with energy gained from renewable energy sources. This not only benefits the environment but can also have a positive financial effect due to the lower prices for renewable energies [3].

However, green energy (solar and wind energy in particular) has the disadvantage of fluctuating yield [3], which may result in an unreliable energy provision. In our work, we focus on making server systems resilient against energy scarcity and power failures by combining commercially available Non-Volatile Memory (NVRAM) with software measures rooted in the Operating System (OS). The goal is to suspend the OS when a power failure is recognised and to resume it as soon as a stable power supply is guaranteed. The mechanism is transparent to user space processes, which will continue their execution when the entire system is resumed, with neither data nor progress lost. We use NVRAM because of its persistent nature, coupled with the byte-addressability, which makes it a suitable replacement for volatile DRAM [5]. However, the NVRAM's storage property comes at the cost of slightly lower performance.

We have already adapted the FreeBSD and Linux kernel to explore a NVRAM-only execution model. So the whole kernel and nearly all allocations are exclusively served from persistent memory [4]. This extends to all user space programs that are loaded into NVRAM as well. With NVRAM as the main memory for our system, most of the system is inherently persistent. The only volatile state that is left are processor registers and caches, as well as devices. The slowdown of a NVRAM-only system increases with the degree of parallelism, ranging from 1.08 to 3.71.

The system was then further extended to support a fast suspend/resume mechanism that can save the processor's and devices' state within 3 s. The so-called suspend-to-NVRAM can be used in versatile ways. Either to quickly react to an unstable supply of power and to avoid data loss due to power outages, or to power down idling servers when the data centre's utilisation is low. The latter point saves energy in general and also extends to other areas within data centres, such as the cooling system. When the systems start up again, the software-managed caches of the previous execution are still available, and the services can continue without a reduced performance during a warm-up phase.

In order to mask the decreased performance of a NVRAM-only system, we plan to selectively mix DRAM with NVRAM, resulting in a NVRAM-mostly system. One measure is capacity scaling, which combines NVRAM and DRAM within the virtual memory management and tries to keep the content of frequently used pages in the faster tier memory. The tiered memory system is about 31.37 % faster for workloads with random memory accesses.The capacity scaling is not limited to NVRAM but can also be used with memory such as High-Bandwidth Memory (HBM), which is even faster than DRAM [2]. Furthermore, even memory built into a remote server should be usable when attached via CXL.mem. The additional volatile memory increases the state that must be saved during the suspend operation. Therefore, the amount of the utilised volatile memory directly influences the duration of a suspend-to-NVRAM.

We are currently improving the system by reducing the suspension time. It is dominated by saving the device's state, which takes up to 94 % of the time to suspend. In addition, the swapping decisions can be improved by incorporating more runtime information into capacity scaling.

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — project number 501993201.

[1] International Energy Agency. Electricity 2024 analysis and forecast to 2026. https://www.iea.org/reports/electricity-2024. Last visited: 2024-06-19.

[2] Giersch et al. Virtual Memory Revisited for Tiered Memory. APSys'24.

[3] Hönig et al. How to Make Profit: Exploiting Fluctuating Electricity Prices with Albatross, a Runtime System for Heterogeneous HPC Clusters. ROSS'18, New York, NY, USA, 2018.

[4] Rabenstein et al. Back to the core-memory age: Running operating systems in NVRAM only. In *Architecture of Computing Systems*, pages 153–167. Springer Nature Switzerland, 2023.

[5] Intel Corp. Achieve greater insight from your data. https://www.intel.com/content/www/us/en/products/docs/memory-storage/optane-persistent-memory/optane-persistent-memory-200-series-brief.html, 2022. Last visited: 2023-11-09.