

Ačkoliv je implementace ovladače blokového zařízení v uživatelském prostoru moderního operačního systému možná, je velmi neobvyklá a často dosahuje velmi nízkého výkonu. Možnost snadno implementovat vysoce výkonný ovladač blokového zařízení v uživatelském prostoru by dovolila rychlejší ověřování vědeckých poznatků z oblasti datových úložišť a umožnila by navrhovat bloková zařízení, která fungují velmi odlišně od těch tradičních.

V této práci představujeme „Block Device in Userspace” (BUSE), což je linuxový modul a komunikační protokol, který umožňuje vývoj vysoce výkonného ovladače blokového zařízení v uživatelském prostoru. V porovnání s dosud existujícími přístupy BUSE výborně škáluje na moderních vícejádrových architekturách, poskytuje nejméně 7x vyšší propustnost a nabízí výrazně jednodušší nastavení. Modul komunikuje s ovladačem v uživatelském prostoru přes sdílenou paměť, což eliminuje nadbytečné kopírování paměti. BUSE dále řeší případné konzistenční problémy typu zápis po zápisu či čtení po zápisu, které jsou způsobeny více frontami bez synchronizace v příslušné části operačního systému. Tím je výrazně usnadněna implementace ovladače v uživatelském prostoru, která se může plně věnovat problémové doméně.

Jako demonstraci použití BUSE práce dále představuje „Block Device in S3” (BS3). Jedná se o vysoce efektivní implementaci blokového zařízení v uživatelském prostoru, která používá Amazon S3 (nebo jiné kompatibilní úložiště) pro ukládání dat. BS3 může být použita jako běžný disk, který poskytuje propustnost více než 10 GB/s. To je více než poskytují nejrychlejší lokálně připojené NVMe SSD přes PCIe 3.0 4x. V porovnání s komerčním řešením CloudBD, které avizuje vysoce výkonný disk přes Amazon S3, BS3 dosahuje až 130x vyšší propustnosti. Odolnost (durability) uložených dat v Amazon S3 je 99,99999999 % (jedenáct devítek), což dělá z BS3 velmi odolné úložiště dat. Navíc je BS3 prefix-konzistentní za všech okolností, díky čemuž zachovává záruky souborových systémů na konzistenci po poruše.