段卓辉

zhduan@hust.edu.cn · 13627101736 · Zhuohui Duan · Projects · Website

教育背景

华中科技大学, 湖北, 武汉 2018 – 至今

博士研究生 计算机软件与理论，预计 2022 年 6 月毕业

华中科技大学, 湖北, 武汉 2016 – 2018

硕士研究生 计算机技术

江汉大学, 湖北, 武汉 2012 – 2016

学士 计算机科学与技术

工作介绍

HME：轻量级异构内存模拟器 2016 – 2018

近年来，新兴的非易失性存储器（NVM）技术已被广泛研究。因为商业 NVM 硬件成本高，研究入门门槛难，这些研究主要依赖于周期精确的架构模拟器。然而，目前的模拟方法要么太慢，要么不能模拟复杂和大规模的工作负载。

- 工作要点
  - 提出了一个轻量级的混合内存仿真器，利用 Intel CPU 的现有硬件特性，用 DRAM 模拟 NVM 的性能特征, 并在基于 NUMA 的 Intel Xeon 处理器上实现了 HME。
  - 使用远程 NUMA 节点上的 DRAM 来模拟更高的延迟和更低带宽的 NVM。
  - 利用 Intel CPU 提供的 DRAM 热控制接口来限制最大内存带宽。
  - 重新设计了内存分配器，明确支持从 DRAM 或 NVM 区域分配内存。
  - 修改了 Linux 内核以识别 NVM 的内存区域，并扩展 Glibc 库以支持 nvm malloc 接口。
- 发表论文
  - **Zhuohui Duan**, Haikun Liu, Xiaofei Liao, Hai Jin. “A Performance Emulator for Non-volatile Memory”, Symposium on Operating Systems Principles Posters (SOSP Posters), 2017.
  - **Zhuohui Duan**, Haikun Liu, Xiaofei Liao, Hai Jin. “HME: A Lightweight Emulator for Hybrid Memory”, Proceeding of 2018 Design, Automation and Test in Europe Conference and Exhibition (DATE), 2018. (Best Paper Award Nominations)
- 开源代码：https://github.com/CGCL-codes/HME
- 项目依托
  - 国家重点研发计划课题《新型内存和存储技术》科技部项目
  - 国家高技术研究发展计划 (863 计划)《内存计算》项目

HiNUMA：NUMA 感知的异构内存系统数据放置和迁移 2017 – 2019

非统一内存访问（NUMA）架构的特点是不同 CPU 节点上的内存访问延迟不对称。由于非易失性存储器（NVM）和 DRAM 之间相对较大的性能差距，由非易失性存储器（NVM）和 DRAM 组成的异构存储器系统进一步使存储器访问延迟多样化。传统的 NUMA 内存管理策略不能有效地管理异构内存，甚至可能损害应用性能。

- 工作要点
  - 提出了 HiNUMA 对传统 NUMA 机制进行异构内存架构的扩展。HiNUMA 提供了新的内存访问接口，以区分不同 NUMA 组中的 NVM 和 DRAM。
  - 提出了低延迟和高带宽两种 NUMA 拓扑结构的混合内存分配策略，分别用于延迟敏感的应用和带宽敏感的应用。
  - 提出了一种用于异构内存系统新的 NUMA 平衡机制（HANB）。它将数据热度和节点间的带宽利用率都考虑到了页面迁移中。所提出的机制在操作系统（OS）层中实现，无需修改硬件和应用程序。
- 发表论文

  - **Zhuohui Duan**, Haikun Liu, Xiaofei Liao, Hai Jin, Wenbin Jiang, Yu Zhang. “HiNUMA: NUMA-aware Data Placement and Migration in Hybrid Memory Systems”, Proceeding of 2019 IEEE 37th International Conference on Computer Design (ICCD), 2019.
- 项目依托
  - 《基于新型体系结构的高时效大数据处理系统》国防科技创新特区项目
  - 《基于混合内存架构的大数据处理系统》国防科技创新特区项目

Gengar：基于 RDMA 的分布式异构内存共享池 2018 – 2020

字节可寻址非易失性存储器（NVM）技术承诺比 DRAM 的密度更高、成本更低。它们已被越来越多地用于数据中心的应用。尽管以前有许多关于在单机中使用 NVM 的研究，但在分布式数据中心环境中最好地利用它仍然存在挑战。

- 工作要点
  - 提出一个支持 RDMA 的分布式异构内存共享池，Gengar。Gengar 利用 RDMA 读/写原语（即单边 RDMA 模型）来实现低延迟的远程内存访问，而不涉及远程服务器的 CPU 和操作系统。提供了一套简单的 API，以促进分布式异构内存共享池的 RDMA 编程。
  - 提出了一个高效的 DRAM 缓冲方案来加速远程 NVM 访问。通过利用客户端的 RDMA 读/写语义设计了一个轻量级的内存访问监控机制。然后，热数据被缓存在（本地或远程）DRAM 缓冲区，以加速远程 NVM 访问。将分布式 DRAM 缓冲区作为写通缓存来实现，只加速 NVM 的读取操作，而 RDMA 写入操作直接在 NVM 上执行，以保证数据的持久性。DRAM 缓冲机制由 Gengar 在后台执行，对客户端应用程序完全透明。
  - 通过重新设计 RDMA 写原语，提出了一种新的 RDMA 写模式。仍然依靠可靠的连接模式来保证无损和无序的数据传输。利用代理机制来等待 RDMA 工作完成（WC）事件，允许应用程序的执行与 RDMA 数据传输相重叠。通过这种方式，RDMA 写模式可以将 RDMA 写操作的网络往返时间（RTT）从应用程序的关键路径中隐藏。
  - 提供简单的 API 来保证分布式异构内存共享池系统中对象共享的数据一致性。利用租赁分配机制来保证元数据的一致性，并利用轻量级的写锁来保证数据/元数据的一致性。
- 发表论文
  - **Zhuohui Duan**, Haikun Liu, Haodi Lu, Xiaofei Liao, Hai Jin, Yu Zhang, Bingsheng He. “Gengar: An RDMA-based Distributed Hybrid Memory Pool”, Proceedings of the 41th IEEE International Conference on Distributed Computing Systems (ICDCS), 2021.
- 开源代码：https://github.com/CGCL-codes/gengar
- 项目依托
  - 国家自然科学基金面上项目《基于 RDMA 的分布式异构内存池系统》
  - 国家自然科学基金面上项目《异构内存系统的动态重构机制研究》

pRDMA：硬件支持的分布式持久性内存的远程持久性保证 2019 – 2021

持久性内存（PM）和远程直接内存访问（RDMA）技术的出现，激发了人们对在数据中心环境中使用它们的兴趣。然而，由于 RDMA 网络接口卡（RNIC）中的缓存不稳定，PM 和 RDMA 的结合对保证远程数据的持久性提出了重大挑战。持久的 RDMA 操作和 RDMA 更新的可见性在远程持久性内存系统中还没有被充分探索。尽管一些远程过程调用（RPC）通过一组 RDMA 操作支持远程数据持久性，但这些基于软件的解决方案需要远程 CPU 干预，并推迟了远程持久性的可见性。

- 工作要点
  - 对以前基于 RDMA 的 RPC 设计以及它们对持久 RDMA 操作效率的影响进行了比较研究。
  - 基于所学到的经验，设计了一套 RNIC 硬件支持的 RDMA 原语，将数据从 RNIC 的易失性缓存冲到 PM。
  - 基于提议的 RDMA 冲刷原语实现了几个持久的 RPC，以支持远程数据持久性和快速故障恢复。由于持久性 RPC 将数据持久化与 RPC 处理解耦，所以远程数据持久化对应用程序来说比传统 RPC 更早可见。这为通过重叠 RDMA 传输和 RPC 处理来提高应用程序的性能提供了大量机会。这也证明在系统崩溃或断电的情况下，在服务器端恢复不完整的 RPC 是有可能的，而不需要从客户端重新获取数据。
  - 对于一些特殊的硬件特性，如数据直接 I/O，以及典型的 RDMA 传输优化，如数据批处理，讨论了它们对远程数据持久性的影响，然后提出了解决方案。
- 发表论文

  - **Zhuohui Duan**, Haodi Lu, Haikun Liu, Xiaofei Liao, Hai Jin, Yu Zhang, Song Wu. “Hardware-supported Remote Persistence for Distributed Persistent Memory”, Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC), 2021.
- 开源代码：https://github.com/Gumi-presentation-by-Dzh/pRDMA
- 项目依托
  - 国家自然科学基金面上项目《异构内存系统的动态重构机制研究》
  - 《基于 RDMA 的持久内存数据一致性技术》华为公司横向合作项目

MioDB：重新审视异构内存系统中的 KV 存储的日志结构化合并 2020 – 2021

实验研究表明，使用非易失性存储器的基于 LSM 树的 KV 存储的性能瓶颈主要来自于（1）跨内存和存储的昂贵的数据序列化/反序列化，以及（2）内存到磁盘的数据 flush 和磁盘上的数据压缩之间的不对称速度。它们可能会因为写停滞和写放大而导致不可预测的性能下降。

- 工作要点
  - 设计了一个基于跳表的 LSM 树，以充分利用可字节寻址的 NVM 进行 KV 存储，并利用单次 flush 来显著减少 DRAM 和 NVM 之间的数据序列化/反序列化的成本。
  - 利用零拷贝合并和滞后合并的组合来减少写入停滞和写放大。
  - 提出了并行合并，在 LSM 树的所有层次上协调数据 flush 和合并，从而进一步减少写停滞，提高查询性能。
- 开源代码：https://github.com/Gumi-presentation-by-Dzh/mioDB

CLIMBER: 防御相变存储器的翻转地址攻击 2020 – 2021

PCM 由于有限的写入耐久度，容易受到恶意攻击。其单元通常表现出巨大的耐用性变化，这是由于” 过程变化”（Process Variation）造成的。耐久性低的单元（即弱单元）在密集写入攻击下可以在几秒钟内被耗尽。为了延长器件的寿命，已经提出了许多 PV 感知的磨损均衡方案。其基本原理是将预测的密集写入分配给具有高耐力的强单元。这些方案假设未来的写入强度分布与前一个区间的预测一致。然而，它们很容易受到” 翻转地址攻击”（Flipping Address Attack）的影响，在这种情况下，恶意程序可以使在两个相邻的区间内写入强度的分布绝对相反，因此密集写入被分配给弱单元。以前的 PV 感知磨损均衡方案在这种攻击下甚至会加剧弱单元的磨损。

- 工作要点
  - 揭示了另一种变种的翻转地址攻击，称为热-冷页面交换攻击。它利用系统中正常应用产生的写强度分布来隐蔽地攻击薄弱单元。它在运行阶段交换热/冷页，以欺骗损耗均衡算法。
  - 提出了一个新的磨损均衡方案 CLIMBER。它动态地改变有害的地址映射，在运行阶段，大量写入的地址被映射到弱小的单元。因此，CLIMBER 可以纠正磨损均衡算法的预测错误，并根据页面热度的阈值，将密集写入的地址重定向到强单元。
  - 提出了一个 Weak Page Randomly Mapping 方案，以保护最弱的 NVM 单元免受 flipping address 攻击。随机地将冷页映射到包含许多弱页的 NVM 弱区，因此攻击者无法向给定的弱页提供密集的写入流量。通过这种方式，可以将密集写入分散到大量的弱页，从而减轻不一致写入攻击的影响。
  - 我们注意到，CLIMBER 可以很容易地集成到以前的 PV-aware wear leveling 方案中，以防御 flipping address attacks，并进一步提高 PCM 的寿命。
- 开源代码：https://github.com/Gumi-presentation-by-Dzh/CLIMBER

RCCT：基于 ReRAM 的 CIM 架构中计算卸载的编译工具 2020 – 2021

内存中计算（CIM）已经成为一种非冯-诺依曼计算机架构，它可以通过直接在新兴的 ReRAM（如 memristor）crossbars 上进行特定领域的计算来解决” 内存墙” 问题。这些基于 ReRAM 的模拟加速器通常与通用处理器合作使用，以加快各种人工神经网络的应用。在这样的异构计算架构中，传统的软件应该被重新设计和改造，以有效地探索全新的 CIM 加速器。

- 工作要点
  - 提出了一个新的编译工具，使遗留程序自动适应 CPU/CIM 异构架构。依赖于 LLVM 编译器基础设施，将应用程序的源代码编译成 LLVM IR。对于没有源代码的二进制可执行文件，利用反编译工具 McSema 来生成程序的 LLVM IR。
  - 从 LLVM IR 的角度识别并定义了几个可以由基于 ReRAM 的加速器加速的典型计算模式。

  - 此编译工具可以识别可加速的计算模式，如矢量矩阵乘法、位图逻辑运算，并通过非常简单的 API 将这些计算自动卸载到 CIM 加速器上。
  - 对于特定领域的应用，认识到并分类了几种可以通过基于 ReRAM 的 crossbars 的计算模式。
  - 从 LLVM IR 的角度抽象和定义了这些模式，以便此编译工具识别它们，并有效地将计算卸载到基于 ReRAM 的的 crossbars 上。
  - 此编译工具可以自动将程序迁移到 CPU/CIM 异构架构上，而不需要修改应用源代码。它还可以通过利用反编译工具透明地转换没有源代码的二进制可执行文件。
- 开源代码：https://github.com/leibo-hust/RCCT

参与项目

- 国家重点研发计划课题《新型内存和存储技术》科技部项目
- 国家高技术研究发展计划 (863 计划)《内存计算》项目
- 《基于新型体系结构的高时效大数据处理系统》国防科技创新特区项目
- 《基于混合内存架构的大数据处理系统》国防科技创新特区项目
- 国家自然科学基金面上项目《基于 RDMA 的分布式异构内存池系统》
- 国家自然科学基金面上项目《异构内存系统的动态重构机制研究》
- 《基于 RDMA 的持久内存数据一致性技术》华为公司横向合作项目

专利著作

- 专利
  - 《基于 NUMA 架构的异构内存分配和迁移方法》
  - 《基于 RDMA 架构的异构内存共享池构建方法》
  - 《基于 NUMA 架构的异构内存模拟方法》
- 软件著作权
  - 《基于 RDMA 的异构内存共享池软件》
  - 《异构内存系统的内存分配器及其模拟环境软件》

获奖情况

华中科技大学三好研究生	2018
华中科学大学优秀硕士毕业生	2018
研究生国家奖学金	2017
华中科技大学三好研究生	2017
华中科技大学知行奖学金	2017
中兴捧月软件挑战赛华南赛区区域优胜奖	2017
华为精英挑战赛三等奖、武长赛区 35 强	2017
江汉大学校长奖学金	2015
本科生国家奖学金	2014

技能

- 研究方向: 计算机体系结构，异构内存系统，分布式系统
- 具体技能：Intel PMU 编程，Linux 内核编程，Linux 内存系统管理，RDMA 系统编程和设计，异构内存索引设计，LSM-based KV 存储系统
- 平台: Linux
- 编程语言: C, C++
- 语言: 英语 - 熟练