****

现代操作系统

总结报告

|  |  |
| --- | --- |
| 学 院 | 大数据学院 |
| 学 号 | 241812019 |
| 学生姓名 | Xxx |
| 任课教师 | 宋虹 |
| 完成时间 | 2024.10.06 |

：面向无服务器架构中微虚拟机的硬件加速快照压缩

24181201x

**摘 要** 随着无服务器计算在云计算领域的快速发展，冷启动问题逐渐成为影响无服务器应用性能的关键因素。为了应对这一问题，本文学习并且总结了Sabre这一基于硬件加速的虚拟机快照压缩与恢复系统。Sabre通过利用Intel内存分析加速器（IAA）进行硬件加速的压缩和解压缩，显著提高了虚拟机快照的恢复速度，减少了冷启动时间。本文对Sabre的设计、实验结果以及潜在问题进行了详细的复盘和总结，并提出了进一步改进方向。针对可能的硬件依赖性和资源调度问题，本文建议未来引入更智能的调度算法和动态预取策略，以提高系统的适应性和效率。总体而言，Sabre在无服务器计算中的应用展示了硬件加速的巨大潜力，为未来的无服务器架构优化提供了新的思路。

**关键词** 无服务器计算，冷启动，虚拟机快照，硬件加速，压缩，内存分析加速器 (IAA)，预取策略

**:** **Hardware-Accelerated Snapshot Compression for Serverless MicroVM**

24181201x

**Abstract** As serverless computing rapidly evolves in the cloud computing domain, the cold start problem has become a critical factor affecting the performance of serverless applications. To address this issue, this review focuses on Sabre, a hardware-accelerated system for virtual machine snapshot compression and recovery. Sabre leverages the Intel In-Memory Analytics Accelerator (IAA) for hardware-accelerated compression and decompression, significantly improving the recovery speed of virtual machine snapshots and reducing cold start times. This paper provides a detailed analysis of Sabre’s design, experimental results, and potential challenges, along with suggestions for further improvements. To address issues such as hardware dependency and resource scheduling, the paper recommends the adoption of intelligent scheduling algorithms and dynamic prefetching strategies to enhance the system’s adaptability and efficiency. Overall, Sabre’s application in serverless computing demonstrates the immense potential of hardware acceleration and offers new insights for optimizing future serverless architectures.

**Key words** Serverless computing, cold start, virtual machine snapshots, hardware acceleration, compression, In-Memory Analytics Accelerator (IAA), prefetching strategies

1. 研究背景与问题描述
   1. 研究背景

无服务器计算（Serverless computing）是一种云计算模型，因其按需分配资源并且高效弹性地处理短生命周期的任务而备受关注。这种计算模式让开发者可以专注于应用逻辑，而无需关心底层的服务器管理和资源分配[1]。然而，无服务器计算中的冷启动问题，即在函数调用时初始化虚拟机或容器的延迟，极大影响了无服务器应用的性能，尤其是当应用需要频繁处理请求时，冷启动带来的延迟可能会显著增加整体响应时间。

为了应对冷启动问题，虚拟机快照技术成为了一个关键解决方案。通过将虚拟机的状态和物理内存保存到快照文件中，并在后续请求中从快照恢复，系统可以跳过完整的虚拟机启动过程，大幅减少冷启动时间。然而，快照文件的大小和预取效率成为制约这一技术进一步优化的瓶颈。传统的快照预取策略可能需要处理大量内存页面，增加了磁盘I/O和恢复时间。同时，现有的软件压缩方法虽然可以减少快照大小，但往往引入了额外的CPU开销，并且解压缩时间较长，反而增加了恢复时间。

在这个背景下，论文提出了一种名为Sabre的硬件加速解决方案，旨在通过利用Intel的内存分析加速器（IAA）[2]~[4]对虚拟机快照进行快速压缩和解压缩，减少冷启动的开销。Sabre通过硬件加速来处理快照页面的压缩和预取任务，将快照恢复时间与磁盘I/O操作重叠，极大地提升了快照预取效率，并且在不影响系统性能的情况下显著缩小了快照文件的大小。这样，不仅减少了冷启动时间，还降低了存储和传输快照的成本。

* 1. 问题描述

冷启动延迟问题：无服务器计算中每次调用函数时，虚拟机需要从头初始化，导致冷启动延迟，特别是在计算密集型或高并发的环境中，这个问题尤为突出。

快照技术的局限：虽然快照技术可以通过保存虚拟机状态来减少冷启动时间，但现有的快照文件通常较大，预取这些文件的开销较高，并且传统的软件压缩方法引入了额外的计算开销。

硬件加速的潜力：现有的研究主要集中在通过软件优化来减少冷启动时间，而硬件加速（如Intel IAA）为进一步优化虚拟机快照的压缩和恢复带来了新的可能性，但尚未被充分利用。

Sabre通过硬件加速的压缩和预取机制解决了这些问题，显著减少了冷启动延迟，并为无服务器计算的性能提升提供了新的途径。

1. 相关研究与方法
   1. 原有技术基础
      1. 快照

快照是一种技术，它允许将虚拟机的状态和客户操作系统的物理内存存储在本地或远程文件系统中的文件中。通常，在虚拟机和应用逻辑（包括所有依赖项）完全初始化并准备好处理请求后创建快照。在虚拟机的下次调用时，管理程序从快照恢复虚拟机的状态和客户内存，而不是从头启动虚拟机。这大大减少了冷启动开销。

在最基本的情况下，快照包含了整个客户物理内存，这称作全局快照。一些管理程序（如Firecracker）还允许脏内存跟踪功能，它只存储管理程序观察到的客户脏页面。快照可以按照应用程序的依赖层次结构进行组织，这称为增量快照。为解决快照文件过大问题，业内采用按需分页来实现内存恢复。但是分页带来的缺页错误又会大大增加冷启动延迟。

* + 1. 预取

减少页面错误的一个方法是预取快照中的页面。工作集（WS）估算可以有效地进行预取。每个快照都包含一个WS文件，存储了后续调用时可能会访问的页面。根据不同的工作集估算技术，WS文件可以以不同的方式构建。例如，在Record-and-Replay（REAP）技术中，作者建议在无服务器架构函数首次调用时记录所有被访问的客户页面，并将它们放入WS文件中。在下次调用时，可以从磁盘预取WS文件，并将WS页面加载到客户内存中以加快下次冷启动的速度。REAP在函数的不同调用之间具有类似工作集的应用中表现良好。但在工作集差异较大的情况下，REAP可能无法提供良好的性能；在这种情况下，预取其他部分脏页面（甚至所有脏页面）可能更有利。

* + 1. 压缩

预取效率都取决于需要从磁盘恢复的内存大小。通过减少快照或WS文件的大小来提高预取技术的效率。这也减少了存储快照所需的磁盘空间。然而，内存恢复发生在虚拟机启动的关键路径上，对于高压缩比的压缩算法，解压缩可能需要很长时间，造成额外的时间开销。

* 1. 本论文提出研究和方法

Nikita Lazarev、Varun Gohil[5][17]等人设计了一种名为Sabre的硬件加速系统，用于加速无服务器微虚拟机的快照压缩和恢复过程。Sabre通过利用Intel的内存分析加速器（IAA），实现了对虚拟机快照的硬件加速压缩和解压缩，显著提高了快照的恢复速度。Sabre还提供了一种优化的内存预取机制，允许用户根据快照文件的稀疏性选择最佳的预取策略，并且能够在磁盘I/O和解压缩之间实现最大程度的重叠，从而减少冷启动延迟。并且通过IAA进行硬件加速，Sabre能够将快照的大小压缩至4.5倍，并显著加速冷启动过程中内存的恢复。在实际应用中，Sabre与AWS Firecracker微虚拟机无缝集成，并通过优化快照压缩和恢复，使无服务器应用的冷启动时间减少了20%-55%。相比于传统的CPU-only快照恢复方案[6][7]~[8]，Sabre在多个无服务器基准测试中展示了显著的性能提升。

1. 本论文方法的具体描述
   1. 设计前置实验和评估

在论文的第三部分，作者通过一系列实验对Intel内存分析加速器（IAA）的压缩和解压缩性能进行了详细评估。

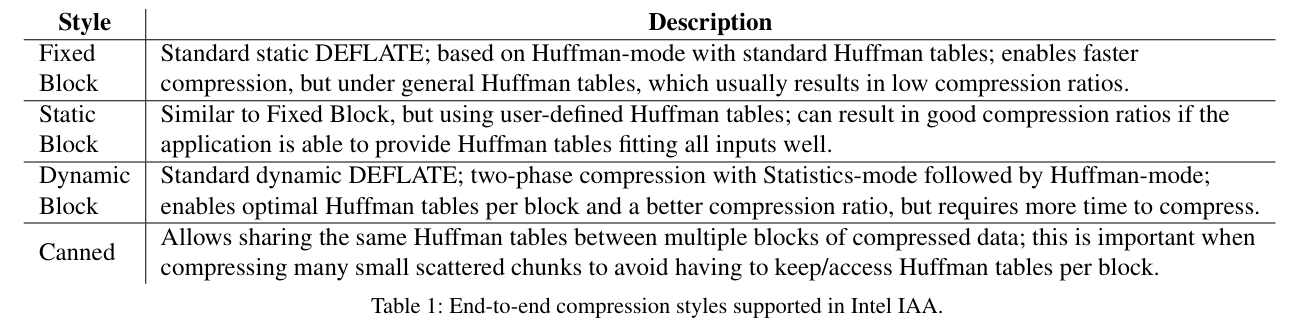


图3.1 IAA支持的几种压缩风格

图3.1显示了IAA支持的集中压缩方式，实验中将根据目标需求和IAA硬件特点选择出Dynamic Block压缩方式[9]~[12]，并采用Silesia压缩数据集以及两个脏内存快照数据集（pythongrpc和pillow）来模拟实际无服务器环境中的工作负载。实验平台基于Intel第四代Xeon可扩展处理器，并利用IAA硬件加速功能进行压缩与解压缩操作[13]。

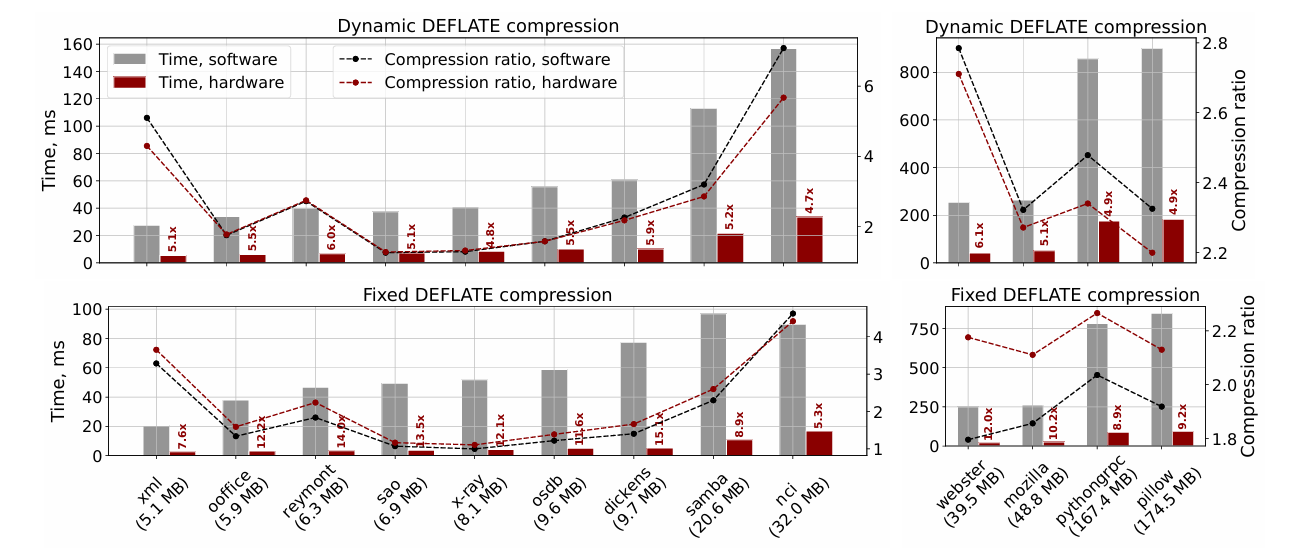


图3.2 实验压缩率和压缩时间显示

图3.2显示了实验结果。实验结果显示，IAA在处理虚拟机快照的压缩任务时显著优于传统的软件压缩方法。具体来说，IAA在压缩脏内存快照时的加速效果明显，压缩比达到了2到4.5倍，且压缩速度比软件实现快6.1到13.5倍。在脏内存快照数据集中的表现尤为突出，压缩加速比达到了9倍。解压缩方面，IAA也表现出显著优势，其解压缩速度比软件实现快一个数量级。

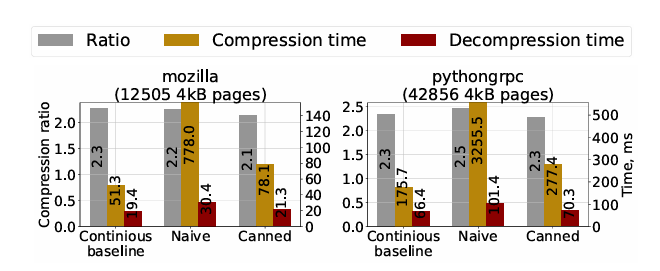


图3.3 页大小为4KB时（解）压缩时间显示

此外，实验还分析了IAA在处理小块分散数据时的性能表现。图3.3显示了Canned和其他方式的结果对比。能发现使用“Canned”操作模式在压缩分散数据块时能够有效减少开销，并使得分散数据块的压缩性能接近处理连续数据块的水平。

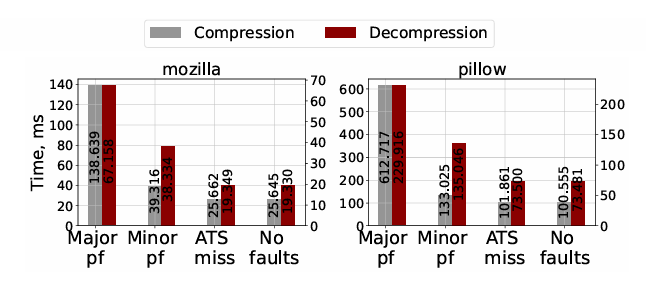


图3.4 使用硬件后效果表现

如图3.4显示，在并行任务处理实验种，IAA在处理多个作业时能够大幅降低延迟，在最大化利用硬件加速的情况下，延迟减少了4到7倍，最高可达17倍。实验发现，主页面错误对性能影响较大，而次页面错误和ATS翻译请求对性能的影响较小。总的来看，通过IAA的硬件加速，Sabre能够有效压缩和解压缩虚拟机快照，显著缩小快照文件大小，同时加速冷启动过程中内存的恢复。

* 1. 设计与实验

Sabre的设计旨在通过硬件加速来优化无服务器微虚拟机的快照压缩和恢复，以减少冷启动的延迟。Sabre利用Intel内存分析加速器（IAA），实现对虚拟机快照页面的高效压缩与解压缩操作。

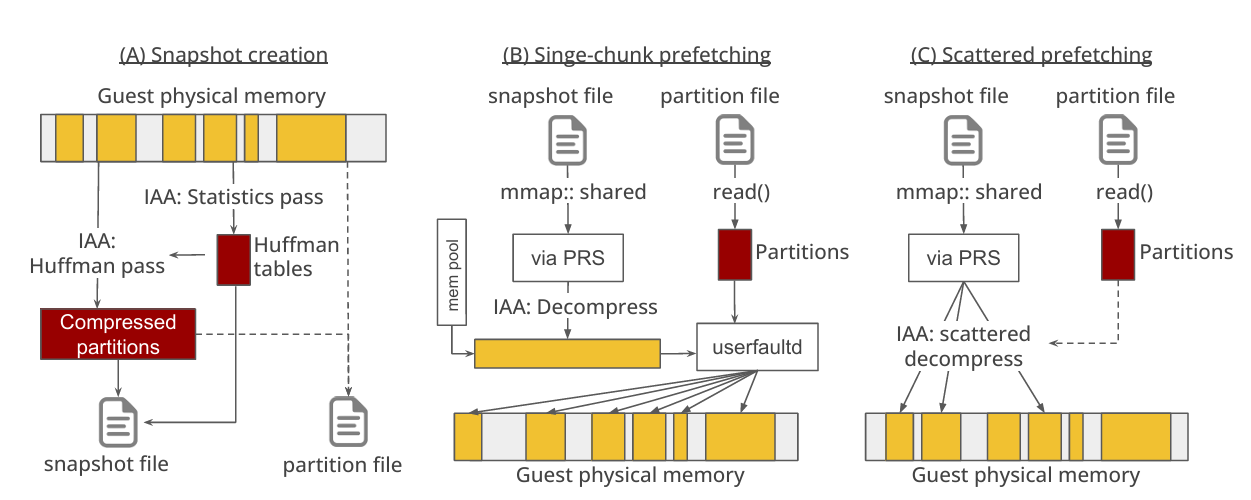


图3.5 Sabre快照的高级综述

如图3.5所示，设计上，Sabre提供了两种内存预取策略：单块预取和分散预取。这两种机制根据快照文件的稀疏度动态选择最优方案，确保压缩和解压缩过程能够与磁盘I/O操作并行执行，从而最大程度减少恢复过程中的延迟。Sabre能够无缝集成现有的虚拟机管理器（如AWS Firecracker），并且兼容多种快照技术，包括脏页面快照和工作集快照。确保压缩和解压缩过程能够与磁盘I/O操作并行执行是这个系统设计的关键，这个并行使得冷启动延迟时间提高显著。

1. 性能评估
   1. 数据包的传输速率

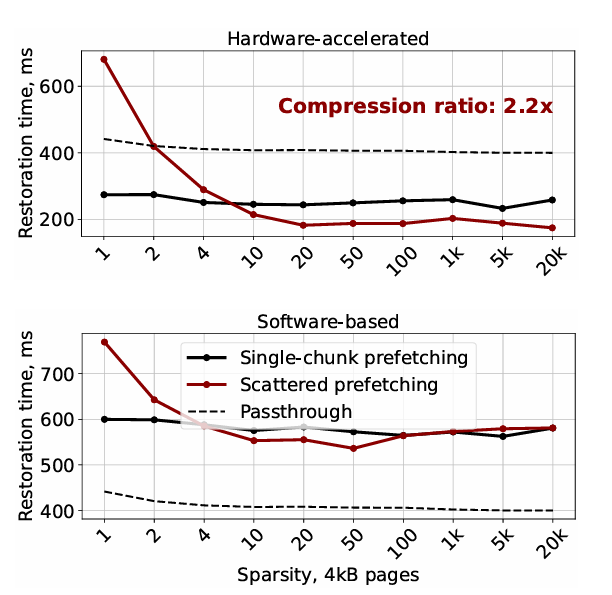


图 4.1 硬件和软件实验结果

Sabre通过多个无服务器应用基准测试对其性能进行了验证，重点评估了差异快照和工作集快照在压缩与恢复过程中的表现。

如图4.1所示，Sabre能够将虚拟机快照的大小压缩至原始大小的4.5倍，并通过硬件加速的内存预取减少冷启动时间。在差异快照场景中，分散预取机制使冷启动延迟减少了最多55%。在工作集快照的场景中，Sabre展示了25%到55%的恢复加速效果，压缩率最高达到4.7倍，显著优于脏页面快照。在内存密集型应用（如python-list基准测试）中，Sabre实现了最高70%的内存恢复速度提升，进一步证明了其在无服务器应用场景中的有效性。

对比硬件和软件的表现，可以发现基于软件加速的表现在基准线之上，而硬件加速的表现则在达到一定稀疏度后有明显改进。这表明，Sabre通过硬件加速成功优化了虚拟机快照的压缩和恢复过程，表现优异。图4.2是整体设计的框架图。

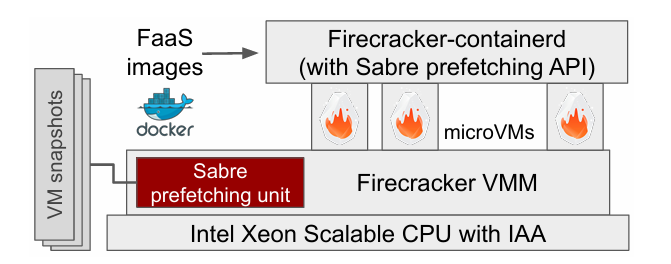


图 4.2 Sabre系统整体设计框架

* 1. 端到端的无服务器架构冷启动时间评估

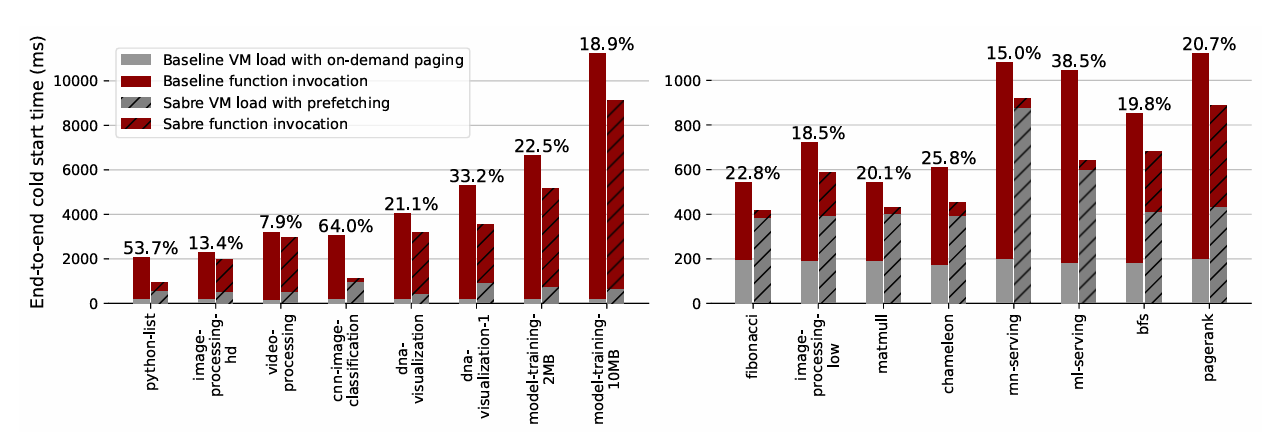


图 4.3 Sabre对Firecracker的默认Diff快照进行端到端的无服务器架构冷启动评估预取

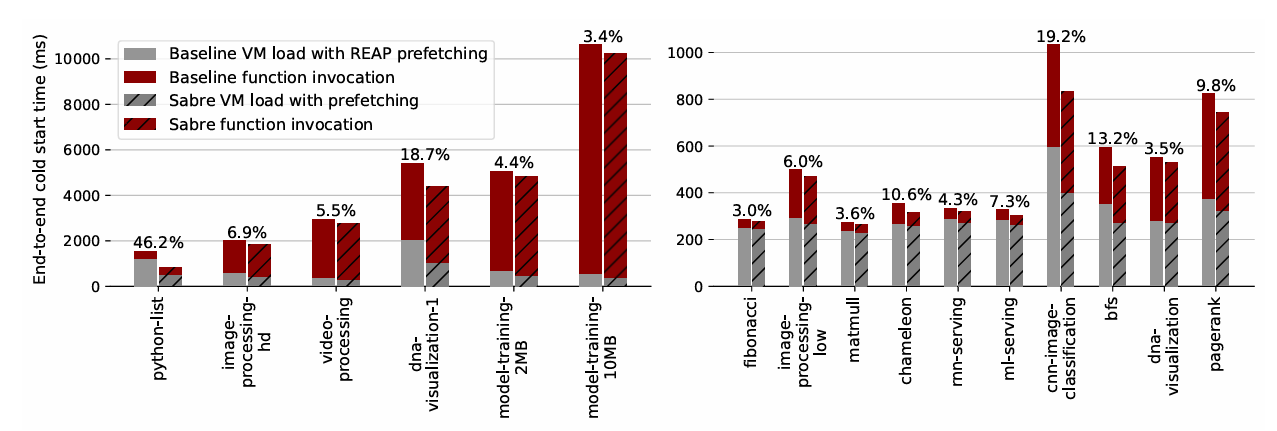


图 4.4 Sabre对REAP快照进行端到端的无服务器架构冷启动评估预取

Sabre进行了端到端的评估，以验证其在实际无服务器应用中的效果。通过多个无服务器基准测试[14][15]对Sabre的冷启动性能和快照压缩效率进行了详细测试。

图4.3和图4.4的实验结果表明，Sabre通过IAA硬件[16]加速实现了显著的压缩和恢复性能提升。在使用差异快照的场景中，Sabre将快照的大小压缩至原始大小的2.5至4倍不等，极大降低了存储和I/O开销。在端到端冷启动延迟上，Sabre减少了冷启动时间，最高可减少60%，特别是在内存密集型应用（如python-list基准）中，Sabre通过快速的内存预取与恢复机制提升了冷启动性能。对于工作集快照，Sabre展现了更高的压缩率，达到4.7倍，并使冷启动延迟减少了多达20%。

表 4.1 Sabre的压缩比和页面预取速度的详细数据

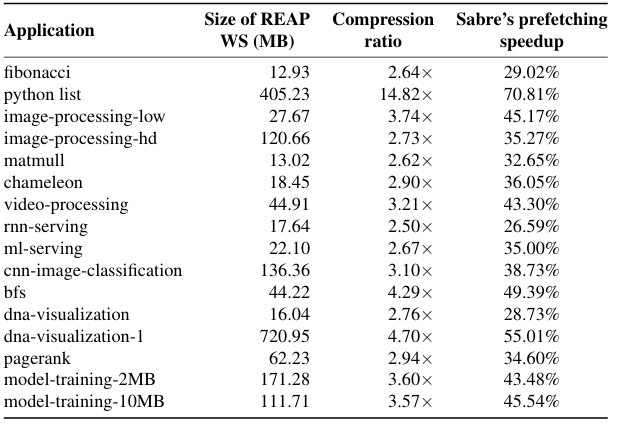


表4.1实现了在所有基准测试中，Sabre的硬件加速压缩显著减少了快照文件的大小，而其内存预取机制在不增加解压缩开销的情况下，提升了快照恢复的速度。尤其是在应用启动延迟受快照大小影响较大的情况下，Sabre通过压缩和预取优化使冷启动时间显著缩短。相比于传统的CPU-only恢复方法，Sabre在大多数场景下的性能都有显著提升。

1. 改进方向
   1. 可能存在的问题

硬件依赖性：Sabre的性能高度依赖于Intel IAA等硬件加速器的可用性。如果系统中没有集成硬件加速器，Sabre的性能提升将受到极大限制。这意味着，在不具备硬件加速能力的云基础设施中，Sabre的效果将大打折扣。

并行负载管理：当前Sabre的内存预取机制主要针对单个虚拟机快照恢复设计，如何在多虚拟机并行运行的场景中，特别是在高负载的情况下，管理和调度硬件资源，可能会成为系统瓶颈。硬件加速器资源的竞争可能会导致性能下降，尤其是在大规模云环境中运行多个无服务器应用时。

快照预取策略：尽管Sabre提供了单块预取和分散预取两种机制，但当快照文件的稀疏性和大小大幅波动时，现有的策略可能无法有效应对。此外，硬件加速器和磁盘I/O的负载不均衡可能导致性能的波动，影响冷启动延迟的稳定性。

兼容性问题：Sabre的当前设计主要基于Firecracker虚拟机管理器，在其他虚拟化技术或环境中的应用和集成（如容器、KVM等）尚未充分验证。因此，在不同虚拟化平台下的兼容性和通用性是一个值得进一步探索的问题。

* 1. 可能的解决方案

针对上述可能存在的问题，以下是一些潜在的解决方案：

增强软硬件适应性：为了降低对硬件加速器的依赖，Sabre可以在没有硬件加速器的环境中引入更多的优化策略，例如通过软件的多线程压缩、并行计算和异步I/O等技术，尽可能缩小与硬件加速之间的性能差距。这样可以提升Sabre在不同硬件环境中的普适性和灵活性。

多虚拟机资源调度：在多虚拟机并行环境下，可以引入智能的资源调度算法，以更高效地分配硬件加速器资源。这种调度机制可以根据虚拟机的优先级、负载大小以及快照文件的紧急程度，动态调整资源分配。同时，可以利用虚拟机的历史运行数据，预测和预调度硬件资源，从而降低资源竞争带来的性能瓶颈。

动态快照预取策略：为应对快照文件大小和稀疏度的波动，可以引入基于机器学习的动态预取策略。通过学习快照文件的访问模式、大小变化规律以及磁盘I/O的负载情况，系统能够自适应地选择最优的预取策略。此外，优化磁盘I/O和解压缩的并行性，以最大限度地提高系统的吞吐量。

广泛兼容性测试：为了提高Sabre在不同虚拟化平台中的兼容性，未来的工作可以扩展对其他虚拟化技术的支持（如KVM、Xen等），并进行广泛的兼容性测试。同时，可以探索将Sabre与容器技术相结合，提供针对容器化应用的压缩与恢复加速方案。

1. 总结与展望

本文提出了Sabre[17]，一种基于硬件加速的无服务器微虚拟机快照压缩与恢复系统。通过利用Intel内存分析加速器（IAA），Sabre成功减少了冷启动时间，并在无服务器架构应用中提升了快照压缩与恢复的效率。实验结果表明，Sabre在多个基准测试中表现出色，尤其是在内存密集型应用和大规模虚拟机快照中，能够显著提升系统性能。

尽管Sabre展示了硬件加速在优化无服务器计算中的潜力，但其硬件依赖性、并行负载管理、快照预取策略以及广泛兼容性等方面仍然存在一些挑战。未来的工作可以进一步改进Sabre的适应性，优化资源调度策略，并扩展其在不同虚拟化技术中的应用。

参考文献

1. Zhuangzhuang Zhou, Yanqi Zhang, and Christina Delimitrou. Aquatope: QoS-and-uncertainty-aware resource management for multi-stage serverless workflows [C]// 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2023). New York, NY, USA: Association for Computing Machinery, 2023: 1-14.I
2. Intel Data Streaming Accelerator (DSA), Architecture [EB/OL]. https://www.intel.com/content/www/us/en/content-details/759709, 2023.
3. Intel Query Processing Library (QPL) [EB/OL]. https://www.intel.com/content/www/us/en/developer/tools/query-processing-library/overview.html, 2023.
4. Introducing Intel Scalable I/O Virtualization [EB/OL]. https://www.intel.com/content/www/us/en/developer/articles/technical/introducing-intel-scalable-io-virtualization.html, 2023..
5. Implementation of Sabre [EB/OL]. Barabanshek/Sabre, 2024. https://github.com/.
6. KVM Live Migration with IAA In-Memory Compression [EB/OL]. https://lore.kernel.org/all/20240319164527.1873891-1-yuan1.liu@intel.com/T/, 2024.
7. Alexandru Agache, Marc Brooker, Alexandra Iordache, Anthony Liguori, Rolf Neugebauer, Phil Piwonka, and Diana-Maria Popa. Firecracker: Lightweight virtualization for serverless applications [C]// 17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20). Santa Clara, CA: USENIX Association, 2020: 419-434.
8. Mahdi Aiash, Glenford Mapp, and Orhan Gemikonakli. Secure live virtual machines migration: Issues and solutions [C]// 2014 28th International Conference on Advanced Information Networking and Applications Workshops. 2014: 160-165.
9. Jacob Breiholz, Farah Yahya, Christopher J. Lukas, Xing Chen, Kevin Leach, David Wentzloff, and Benton H. Calhoun. A 4.4 nw lossless sensor data compression accelerator for 2.9x system power reduction in wireless body sensors [C]// 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS). 2017: 1041-1044.
10. Anjali, Tyler Caraza-Harter, and Michael M. Swift. Blending containers and virtual machines: A study of Firecracker and gVisor [C]// Proceedings of the 16th ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments (VEE '20). New York, NY, USA: Association for Computing Machinery, 2020: 101-113.
11. Joao Carreira, Pedro Fonseca, Alexey Tumanov, Andrew Zhang, and Randy Katz. Cirrus: A serverless framework for end-to-end ML workflows [C]// Proceedings of the ACM Symposium on Cloud Computing (SoCC '19). New York, NY, USA: Association for Computing Machinery.
12. Lixiang Ao, George Porter, and Geoffrey M. Voelker. Faasnap: Faas made fast using snapshot-based VMs [C]// Proceedings of the Seventeenth European Conference on Computer Systems (EuroSys '22). New York, NY, USA: Association for Computing Machinery, 2022: 730-746.
13. Jianyu Chen, Maurice Daverveldt, and Zaid Al-Ars. FPGA acceleration of Zstd compression algorithm [C]// 2021 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). 2021: 188-191.
14. Amine Barrak, Fabio Petrillo, and Fehmi Jaafar. Serverless on machine learning: A systematic mapping study [J]. IEEE Access, 10:99337-99352, 2022.
15. Ana Klimovic, Yawen Wang, Patrick Stuedi, Animesh Trivedi, Jonas Pfefferle, and Christos Kozyrakis. Pocket: Elastic ephemeral storage for serverless analytics [C]// 13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18). Carlsbad, CA: USENIX Association, 2018: 427-444.
16. Luiz Barroso and Urs Hoelzle. The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines [M]. MC Publishers, 2009.Sabre IAA Benchmarks [EB/OL]. https://github.com/barabanshek/IAA\_benchmarking.git, 2024.
17. Zhuangzhuang Zhou, Varun Gohil, James Tsai, Andy Anderson, Bhushan Chitlur, Zhiru Zhang, and Christina Delimitrou. Sabre: Hardware-Accelerated Snapshot Compression for Serverless MicroVMs [C]// 18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 2024). 2024: 1-14.