arXiv:2304.11921v1 [cs.DC] 24 Apr 2023

新一代 SX-Aurora TSUBASA 矢量超级计算机的性能评估

Keichi Takahashi[1 [0000−0002−1607−5694]]、Soya Fujimoto[2], Satoru Nagase[2], Yoko Isobe[2], Yoichi Shimomura[1], Ryusuke Egawa[3 [0000−0001−8966−867X]]和 Hiroyuki Takizawa[1 [0000−0003−2858−3140]]

[1] 东北大学
{keichi,shimomura32,takizawa}@tohoku.ac.jp[2] NEC公司
{s-fujimoto,s.nagase,y-isobe-pi}@nec.com[3] 东京电机大学 egawa@mail.dendai.ac.jp

摘要。 数据移动是现代高性能计算系统在性能和能效方面的关键瓶颈。长期以来，NEC SX系列超级计算机通过为应用提供充足的内存带宽，加速了内存密集型高性能计算应用。本文分析了一台搭载全新矢量引擎（VE30）处理器的 SX-Aurora TSUBASA 原型超级计算机的性能。VE30 是矢量引擎处理器系列的首次重大升级，其更新的内存子系统显著提升了内存访问性能。此外，它还引入了新指令，并采用了专为加速内存密集型应用而设计的架构改进。通过标准基准测试，我们证明，在内存密集型应用的性能和效率方面，VE30 均显著优于其他处理器。我们还使用包括 SPEChpc 在内的应用对 VE30 进行了评估，并表明 VE30 能够以高性能运行实际应用。最后，我们讨论了充分发挥 VE30 性能的调优技术。

关键词： 性能评估· SX-Aurora TSUBASA · 内存密集型应用 · 向量处理器 · 向量超级计算机

1 引言

内存墙是高性能计算领域长期面临的一项挑战，指的是计算系统中算术计算性能与内存性能之间的差距不断扩大。由于内存墙问题，内存密集型应用受数据移动所限，无法充分利用系统的算术计算性能。这不仅会损害应用的性能，还会降低能效。因此，高性能计算领域一直在积极探索应对内存墙的新型架构，例如采用高带宽内存设备作为片外内存 [3, 21], , 以及实现大容量片上

内存 [13, 18], , 并通过在处理单元之间直接交换数据来减少内存访问 [8, 9]。

然而，这些异构架构与通用 CPU 完全不同，并且通常要求程序员熟悉非常规的编程模型。例如，设备可能要求比 CPU 高出多个数量级的并行度，或者内存子系统中的数据移动可能需要由程序员进行显式管理。因此，为此类新兴硬件开发软件通常既耗时又昂贵。如果拥有大量用户，为专用系统开发优化软件的成本可以得到分摊（例如，深度学习）。然而，在高性能计算领域，科学上十分重要的软件包往往仅由少数几名程序员维护，并被一小群用户使用。在这种情况下，开发成本将变得难以承受。

NEC 的 SX-Aurora TSUBASA（SX-AT）超级计算机通过将最新的内存技术与向量架构独特结合，旨在同时实现世界级的内存性能和高生产力。向量架构拥有悠久的历史，近年来又重新受到业界关注。这一趋势可从 ARM 可扩展向量扩展 [25] 以及即将推出的 RISC-V 向量扩展 [17], 中看出；二者都深受向量架构的启发。由于大多数高性能计算应用都表现出较高的数据级并行性，而向量化编译器可以自动利用这种并行性，因此，面向通用 CPU 的传统软件只需进行少量修改即可运行。为了持续向高性能向量核心提供数据，高带宽内存（HBM）与处理器紧密耦合。因此，SX-AT 在确保程序员生产力的同时，为应用提供了极高的内存性能。

NEC 最近一直在为 SX-AT 试制一种名为 Vector Engine 3.0（VE30）的全新矢量处理器。VE30 相较于上一代 Vector Engine 系列实现了大幅跃升，带来了超越峰值计算性能和内存性能提升的多项架构改进。具体而言，VE30 在内存层次结构中引入了可绕过的每核私有 L3 缓存这一新层级，以加速缓存密集型应用。此外，VE30 还新增了一条指令，可在具备计算能力的末级缓存中执行间接寻址向量累加。

由于这些改进的综合效果有望使应用获得超越峰值性能提升的加速，因此无法简单估算应用性能。为此，我们对一台基于 VE30 处理器的下一代矢量超级计算机进行了首次性能分析。本文的主要贡献总结如下。

– 这是首项评估配备 VE30 处理器的下一代矢量超级计算机性能的研究。我们采用业界标准基准测试和多个应用，评估 VE30 的实际性能以及基本性能。– 本文使用微基准测试分析了 VE30 新引入的各项架构改进所带来的性能增益。– 本文讨论了如何利用 VE30 的新架构能力，通过性能调优技术加速应用性能。

本文其余部分的组织结构如下。第 2 节介绍 NEC SX-AT 超级计算机，并描述 VE30 处理器的基本架构。

第 3 节使用标准基准测试、微基准测试和实际工作负载，对 VE30 的性能进行了全面评估。第 4 节讨论性能调优技术，以充分发挥 VE30 处理器的潜力。第 5 节总结全文。

2 SX-Aurora TSUBASA VE30 概述

本节首先概述 SX-AT 超级计算机的架构，并介绍新开发的 VE30 处理器。随后，我们将描述 VE30 相较于前代产品的架构增强。

2.1 SX-Aurora TSUBASA 产品系列

SX-Aurora TSUBASA (SX-AT) 是 NEC SX 矢量超级计算机系列的最新产品系列。SX-AT 在继承前代产品成熟且成功的设计理念的同时，也采用了当前的事实标准 高性能计算软件生态系统。第一代 SX-AT 基于 Vector Engine 1.0 (VE10) 处理器，于 2018 [11], 发布；随后，第二代 SX-AT 基于 Vector Engine 2.0 (VE20) 处理器，于 2020 [6]发布。本文评估的第三代 SX-AT 基于 Vector Engine 3.0 (VE30) 处理器，目前仍在开发中，预计近期发布。

SX-AT 采用异构架构，由向量主机 *(VH)* 和矢量引擎 *(VE)*组成。VH 是一台 x86 服务器，负责运行操作系统并执行进程管理、内存管理和 I/O 等任务。VE 是一种在 PCI Express (PCIe) 卡上实现的向量处理器，用于执行应用程序。VH 通过 PCIe 链路与 VE 通信，并对其进行控制。

虽然从表面上看，VE 类似于 GPU 等加速器，但其执行模型与传统加速器有很大不同。应用程序完全在 VE 上执行，系统调用则转发到 VH，由运行在 VH 上的代理进程处理。这种设计消除了内核启动开销，并减少了传统加速器中存在的数据传输。此外，这种设计使用户能够使用标准的基于 MPI 和 OpenMP 的编程模型开发应用程序，无需了解特定于供应商的编程语言或框架。

2.2 VE30 处理器的基本架构

图 1 展示了 VE30 处理器原型的概况，图 2 描绘了 VE30 处理器的内存层次结构。VE30 处理器集成了 16 个向量核心、一个共享末级缓存和六个 HBM2E 模块。每个向量核心的性能最高可达 307.2 GFLOP/s（双精度）或 614.4 GFLOP/s（单精度），因此单个插槽的总性能可达 4.91 TFLOP/s（双精度）或 9.83 TFLOP/s（单精度）。六个 HBM2E 模块的容量为 96 GB，可为核心提供总计 2.45 TB/s 的内存带宽。共享末级缓存的容量为 64 MB。核心与末级缓存通过二维片上网络互连

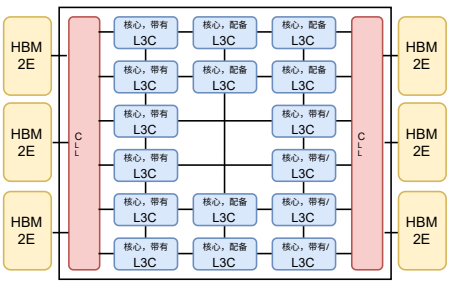


图 1：VE30 处理器的框图。

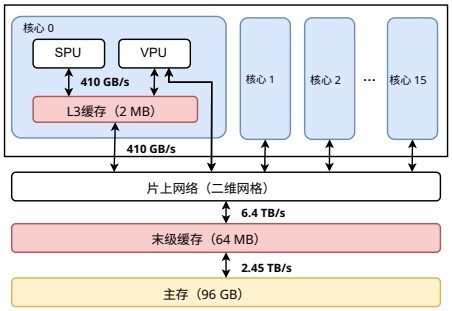


图 2：VE30 处理器的内存层次结构。

（片上网络）。VE30 提供一种分区模式，可将处理器中的核心、末级缓存和高带宽内存划分为两个非一致性内存访问节点。这通过缓解片上网络中的拥塞，提高了总有效LLC带宽，并有利于 LLC 密集型应用。

VE30 中的一个向量核心由标量处理单元（SPU）和向量处理单元（VPU）组成。SPU 获取并解码指令、执行标量指令，并将向量指令分派给 VPU。一个 SPU 包含 64 KB 的 L1 指令缓存、64KB 的 L1 数据缓存和 512 KB 的统一 L2 缓存。一个 VPU 包含 64 个架构向量寄存器，这些寄存器会重命名为 188 个物理向量寄存器。单个向量寄存器最多可容纳 256 个双精度浮点元素（即2 KB）。一个 VPU 包含 32 条向量流水线，每条流水线具有 3 个融合乘加（FMA）执行单元。因此，一个向量核心总共可在单个周期内执行 96 次 FMA 运算。SPU 和 VPU 共享一个 2 MB 的统一 L3缓存。

2.3 VE20 处理器的架构改进

与其前代产品相比，VE30 的内存子系统有了显著提升。首先，内存层次结构中引入了新的层级，即每核私有 L3 缓存，这缓解了末级缓存争用，使缓存密集型应用能够获得更高的性能。其次，末级缓存的容量和带宽分别增加了 4× 和 2.13×。第三，高带宽内存的容量和带宽也均有所提升。峰值高带宽内存带宽从 1.53 TB/s 提升至 2.45 TB/s，增加了 1.60×；高带宽内存容量则从 48 GB 翻倍至 96GB。这些对内存子系统的大幅改进预计将显著加速内存密集型应用和缓存密集型应用。

除了对内存子系统进行改进外，核心数量也从 10 个增加到 16 个，使峰值单插槽性能从 3.07 太浮点运算每秒提升至 4.91 太浮点运算每秒。值得注意的是，尽管核心数量有所增加，每核缓存和内存性能也都得到了提升或保持不变。

此外，还对核心进行了多项改进。首先，VE30 放宽了对单精度浮点向量的对齐要求，并提升了

单精度应用的性能。其次，VE30 引入了专用硬件机制，用于加速采用间接寻址的向量累加。这些改进不会直接提升峰值 FLOP/s 速率，但预计将有助于提升实际应用的性能。

3 性能评估

本节首先使用业界标准基准测试揭示 VE30 的基本性能。随后，我们通过微基准测试考察 VE30 引入的架构改进所带来的性能提升。最后，我们使用代表实际应用的工作负载评估 VE30 在真实场景中的性能。请注意，VE30 的性能测量是在原型软件和硬件上进行的，因此，最终产品上的结果可能会有所变化。

3.1 评估环境

表1：评估处理器的规格。

	VE Type 20B	VE Type 30A	A64FX	Xeon Platinum 8368	A100 80 GB PCIe
频率 [GHz]	1.6	1.6	2.2	2.4	1.412
性能每核心 [GFLOP/s]	307（双精度） 614（单精度）	307（双精度） 614（单精度）	70（双精度） 140（单精度）	83.2（双精度）[1] 166（单精度）	181（双精度），带张量核心 90（双精度），不带张量核心 181（单精度）
核心数量	8	16	48	38	108
性能每插槽 [太浮点运算每秒]	2.4（双精度） 4.9（单精度）	4.9（双精度） 9.8（单精度）	3.3（双精度） 6.7（单精度）	3.1（双精度）[1] 6.3（单精度）	19.5（双精度），使用张量核心 9.7（双精度），不使用张量核心 19.5（单精度）
末级缓存带宽 [TB/s]	3.0	6.4	3.6	3.2[2]	4.9[2]
末级缓存容量 [MB]	16	64	32	54	40
内存带宽 [TB/s]	1.53	2.45	1.024	0.204	1.935
内存容量 [GB]	48	96	32	256	80
制程工艺 [nm]	16	7	7	10	7

表1总结了本次评估中所使用处理器的规格。我们将 VE30 与高性能计算中使用的多种最新处理器进行比较，涵盖向量处理器、GPU、众核处理器和通用 CPU：NEC Vector Engine Type 20B（VE20 的 8 核 SKU） [6], NVIDIA A100 40 GB 和 80GB PCIe 型号 [3], Fujitsu A64FX [21], 以及 Intel Xeon Platinum 8368（IceLake-SP） [16]。如表1所示，纳入张量核心后，A100 的峰值性能翻倍。在计算时，我们使用包含张量核心的峰值性能直到

[1] 峰值性能根据所有核心均处于活动状态时的 AVX-512 Turbo 频率计算。由于制造商未公开峰值带宽，因此使用经验屋顶线工具包 (https://bitbucket.org/berkeleylab/cs-roofline-toolkit)测量末级缓存带宽（IceLake-SP 上为 L2 带宽）。[2]

HPL 的效率，以及其他基准测试不包含张量核心时的峰值性能。这是因为除 HPL 外，所有基准测试都不使用张量核心。

VE30 的多节点测量在一个由 16 个 VH 组成的集群上进行，这些 VH 通过双轨 InfiniBand HDR 网络互连。每个 VH 配备 8 张 Vector Engine Type 30A 卡、一个 AMD EPYC 7713P 处理器和 512 GB DDR4-3200 SDRAM。

3.2 基本基准测试

我们使用 HPC 中四种广泛认可的基准测试来评估 VE30 的基本性能：High Performance Linpack (HPL) [5] 基准测试、STREAM 基准测试、High Performance Conjugate Gradients (HPCG) [4]以及 Himeno 基准测试 [7]。

HPL 是一种计算密集型基准测试，使用带部分选主元的 LU 分解求解稠密线性方程组。STREAM 基准测试用于测量有效内存带宽。HPCG 是一种内存密集型基准测试，使用共轭梯度法和几何多重网格预条件器求解稀疏线性系统。Himeno 基准测试同样是内存密集型基准测试，使用雅可比方法求解泊松方程。只有 Himeno 基准测试使用单精度浮点数进行计算，其余基准测试均使用双精度浮点数。由于我们无法获得针对 A64FX 处理器优化的 HPL 和 HPCG 可执行文件，因此 A64FX 的 HPL 和 HPCG 性能是根据基于 A64FX 的系统在 Top500 上的结果计算得出的（*Fugaku* [21]）。

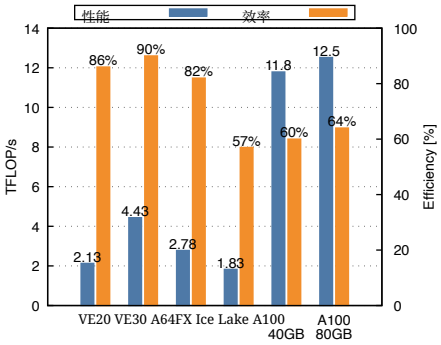


图 3：HPL 基准测试性能。

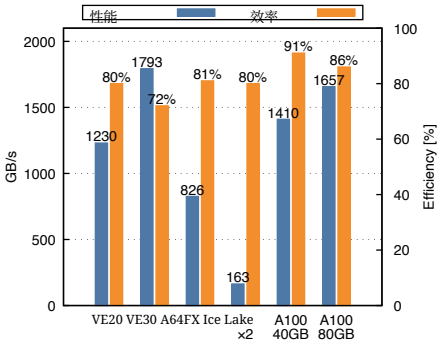


图 4：有效内存带宽。

图 3 比较了不同处理器的 HPL 性能。NVIDIA A100 明显优于其他处理器。A100 40 GB 型号的 HPL 性能超过 11.8 TFLOP/s，而 80 GB 型号由于问题规模更大且 TDP 更高，性能略高，达到 12.5 TFLOP/s。VE30 处理器的性能为 4.43 TFLOP/s，超过了 A64FX 和 IceLake-SP。在效率方面，VE30 以 90% 的效率位居首位，其次是 A64FX 和 VE20。由于功耗限制，A100 无法维持 GPU 加速时钟频率，因此效率相对较低。

图 4 比较了使用 STREAM 测得的不同处理器的有效内存带宽。VE30 的有效内存带宽超过 1.79TB/s，在评估处理器中明显最高。与上一代 VE20 相比，VE30 的有效内存带宽高出 1.45× 。

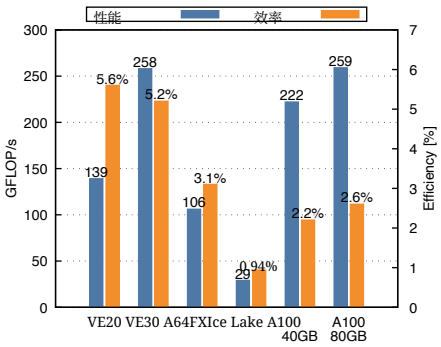


图 5：HPCG 基准测试性能。

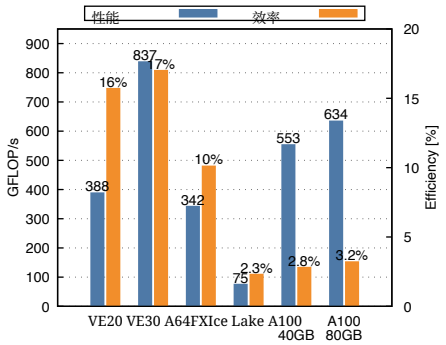


图 6：Himeno 基准测试性能（XL 规模）。

图 5 展示了所评估处理器的 HPCG 性能。VE30 达到 258 GFLOP/s，性能优于 VE20、A64FX、IceLake-SP 和 A100 40 GB 型号。其性能与 A100 80 GB 型号几乎相同。在效率方面，VE30 达到了峰值性能的 5.2%，显著高于其他处理器：比 A100 80GB 型号高 1.97× ，比 IceLake-SP 高 5.72×。此外，VE30 在所有处理器中能效最高。VE30 执行 HPCG 时的能效达到 1.034 GFLOP/s/W，而 A100 40 GB 和 80 GB 型号分别达到 0.909 GFLOP/s/W 和 0.999 GFLOP/s/W。这些结果表明，VE30 成功兼顾了内存性能与浮点性能，而其他处理器则过度侧重浮点性能，忽视了内存性能。

图 6 展示了 Himeno 基准测试的性能。VE30 是所有处理器中性能最佳的型号，达到 837 GFLOP/s，分别是 A100 40 GB 和 A100 80 GB 型号的 1.51× 倍和 1.32× 倍。有趣的是，加速比超过了内存带宽的差异。例如，VE30 的内存带宽比 A100 80 GB 型号高 1.27×，但其性能却高出 1.32×。相较于达到 388GFLOP/s 的 VE20，其加速比为 2.15×，同样远高于峰值内存带宽提升的 1.60× 。这可能是因为 VE30 放宽了对单精度向量的对齐限制，从而能够更高效地执行单精度应用。

最后，我们评估 HPL、HPCG 和 Himeno 基准测试的多节点可扩展性。图 7 展示了两个基准测试的多节点性能随 VE 数量变化的情况。结果表明，三个基准测试从 1 个 VE 扩展到 128 个 VE 时几乎呈线性扩展，效率仅略有下降。在 128 个 VE（即 16 个 VH）上，HPL 性能达到 537 太浮点运算每秒，效率为 85.5%。HPCG 基准测试

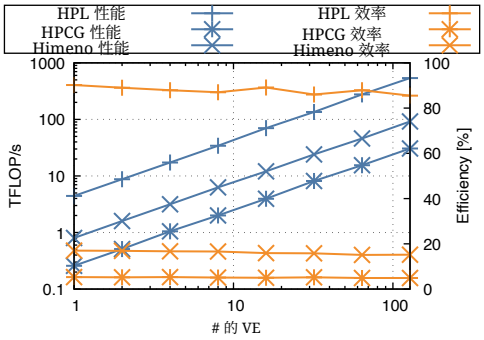


图7：HPL、HPCG 和 Himeno 基准测试的多节点扩展性能。

表2：东北大学内核集合概览。

内核	领域	瓶颈
Earthquake [1]	地震学	内存带宽
湍流 [27]	流体动力学	末级缓存带宽
Antenna [19]	电子学	内存带宽
Land Mine [20]	电子学	内存带宽
涡轮机 [27]	流体动力学	内存延迟
等离子体 [10]	地球物理学	内存延迟

在 128 个 VE 上达到 30.6TFLOP/s，效率为 4.9%。Himeno 基准测试在 128 个 VE 上达到 919 TFLOP/s，效率为 15.2%。

3.3 架构改进评估

可绕过的 L3 缓存 VE30 将每核私有 L3 缓存纳入内存层次结构。此设计选择基于以下观察：在 VE 的前几代产品上，缓存密集型应用的末级缓存性能有所下降。这主要是由于片上网络拥塞以及末级缓存中的缓存竞争。引入私有 L3 缓存有望通过缓解片上网络拥塞和末级缓存竞争，提高有效缓存带宽。

此外，软件可以绕过 L3 缓存。与中央处理器和图形处理器中的非时序加载和存储类似 [14], ，每条加载或存储指令都可以指定是否绕过 L3 缓存。对具有高时间局部性的数据进行选择性缓存，有望减少缓存污染，并使应用能够有效利用有限的缓存容量。从程序员的角度来看，只需在源代码中插入编译器指令 #pragma _NEC on_adb(var) 即可启用选择性缓存，其中 var 表示要缓存在 L3 中的数组。

请注意，L3缓存旁路与前几代 VE 中提供的末级缓存保留控制 [15]有所不同。末级缓存保留控制允许应用程序在发出加载和存储操作时，将数据标记为时间性或非时间性。随后，末级缓存会在驱逐缓存行时优先保留时间性数据，而非时间性数据。不过，即使某次访问被标记为非时间性，数据仍会缓存在末级缓存中。因此，非时间性数据仍可能占用一定的缓存空间。另一方面，L3缓存旁路则会完全绕过L3缓存。

为了评估 L3缓存 对应用性能的贡献，我们利用 L3缓存旁路功能，比较启用和未启用 L3缓存时应用的性能。这里，我们使用东北大学内核集合 [11, 23], ，这是一组从东北大学网络科学中心用户开发的生产应用中提取的计算内核。如表2所示，

该内核集合包含六个内核，涵盖了种类繁多的科学领域和性能特征。

图 8 展示了启用和未启用 L3 缓存时各内核的性能。结果表明，湍流、Antenna、涡轮机和等离子体明显受益于 L3 缓存。由于 L3 缓存通过处理部分内存请求，节省了末级缓存和内存带宽，因此湍流和 Antenna 等末级缓存密集型和内存密集型应用的运行速度得以提升。相比之下，Earthquake 和 Land Mine 并未从 L3 缓存中受益。这两个内核属于内存密集型，可能数据局部性较差，或者工作集大小过大，无法装入 L3 缓存。对内存延迟敏感的内核涡轮机和等离子体也得到了加速，因为 L3 缓存降低了内存延迟。访问末级缓存的延迟高于访问 L3 缓存，因为末级缓存在物理位置上距离更远，且需要通过可能拥塞的片上网络进行通信。

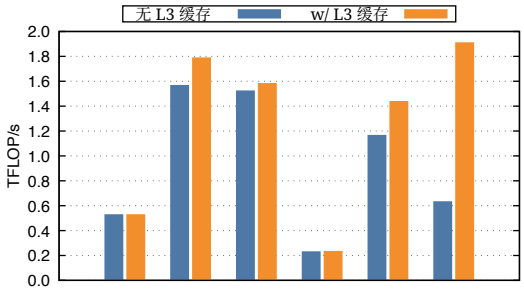


图 8：L3缓存对东北大学内核集合性能的影响。

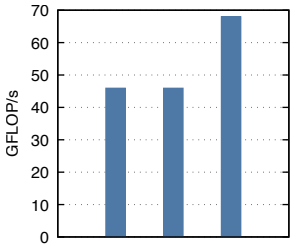


图 9：单精度 27点模板内核的单核心性能。

打包指令的对齐限制放宽 SX-AT 中的打包 指令对 32 位值的向量进行操作，其中向量寄存器的每个 64 位元素都包含一对 32 位值。例如，采用打包格式时，一个向量寄存器可以存储 512 个单精度浮点数。然而，之前几代 VE 施加了对齐限制，要求打包向量的起始地址按 8 字节对齐。否则，无法使用打包格式，并且向量寄存器的每个元素只能包含一个 32 位值，而不是两个。VE30 解除了这一限制，单精度向量只需按 4 字节对齐。

为评估打包格式带来的加速比，我们测量了单精度27点模板内核的性能。图9展示了在VE20和VE30上使用及不使用打包格式时，27点模板内核的单核性能。在VE30上使用打包格式，相较于VE20和未使用打包格式的VE30，性能提升了1.48× 。

索引向量累加的硬件支持VE30处理器引入了对间接寻址向量累加的硬件支持（即稀疏 BLAS 中的 axpyi）。这类计算是有限元和粒子方法等应用的基础。清单1.1展示了索引向量累加的一个示例。在该示例中，数组 y 使用数组 l[i] 作为索引进行间接访问。由于当 l 中的某些索引重叠时会存在循环携带依赖，因此编译器无法自动将此循环向量化。

列表 1.1：索引向量累加。

```
对于 (整数 i = 0; i  <n;i++)
```

在 VE30 之前，程序员需要手动检查 l[i] 是否可能重叠，并在循环中插入 ivdep 或 list_vector 编译器指令。若索引不存在重叠，则指定 ivdep 指令，该指令会直接将循环向量化。若索引确实存在重叠，则必须指定 list_vector 指令。list_vector 指令会指示编译器生成代码，以便：(1) 忽略循环携带依赖，使用向量指令计算结果；(2) 检查索引是否重叠；(3) 使用标量指令修正重叠索引对应的结果。然而，随着向量 l 中重叠索引数量增加，修正所产生的开销也会增加。

VE30 在末级缓存（LLC）中加入了用于原子累加的专用硬件，并新增了一条指令 vlfa。vlfa 指令会将索引向量（列表 1.1 中的 l）和待累加向量（x）发送到末级缓存，然后在末级缓存中执行累加。vlfa 指令的性能应优于 list_vector，因为它无需基于标量的修正，也消除了向量聚集的延迟。此外，程序员无需再费力判断索引是否可能重叠，因此编程效率也得到提升。不过请注意，当重叠索引数量增加时，vlfa 的速度仍会变慢，因为末级缓存中可能发生竞争。

为了研究 vlfa指令的性能，我们使用列表 1.1 所示的索引向量累加内核。在此，我们比较以下五个变体：VE20 和 VE30 上的纯标量版本、VE20 和 VE30 上的 list_vector 版本，以及 VE30 上的 vlfa 版本。我们改变重叠索引的数量，以量化索引重叠导致的性能下降。具体做法是将索引向量 l 初始化如下：

$$l[i] = \begin{cases} 0 & \text{if } i \mod 32 < k \\ i & \text{otherwise,} \end{cases} \tag{1}$$

其中， k 的取值范围为 1 到 32。

图 10 展示了各个变体的索引向量累加性能。显然，当重叠索引数量较少时，vlfa 的性能优于所有其他变体；当 32 个索引中有 4 个重叠时，其速度比 list_vector 快 3.48× 倍，比仅使用标量的方式快 4.72× 倍。重叠索引超过 8 个后，vlfa 的性能开始下降；相同索引超过 20 个时，其性能则低于仅使用标量的方式。不过，在实际应用中，这种高度的地址重叠不太可能出现。这

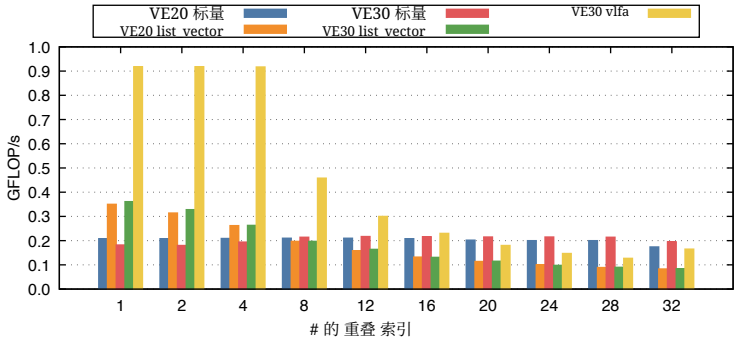


图 10：索引向量累加的单核性能。

清单 1.2：涉及索引向量累加的内核循环。

```
执行 N = nstart,nend如果(flag3(N)==1) 那么COF(7,WI(N),WJ(N),WK(N))=
```

这表明，在 VE30 上，程序员通常无需指定 ivdep 或 list_vector 指令，因此提高了生产力。

清单 1.2 展示了一个出现在实际流体动力学应用中的内核循环。这里，两个三维数组 COF 和 SOC 在单个循环中进行累加。在 VE30 之前，VE 编译器无法将此类代码向量化。在 VE30 上，编译器可以借助 vlfa 指令将此代码向量化。因此，该内核在没有 vlfa 指令的情况下需要 175.6 秒才能在 VE30 上运行，而使用 vlfa 指令时仅需 12.0 秒，加速比达到 14.6× 。

3.4 实际工作负载

SPEChpc 2021 基准测试套件 *SPEChpc* [24] 是由标准性能评估公司（SPEC）开发的基准测试套件，包含一组经过精心选择的应用，代表了广泛的实际高性能计算应用。SPEChpc 基准测试套件的最新版本 SPEChpc 2021 于 2021年10月 发布。它支持多种编程模型，可在中央处理器和图形处理器上运行。在本次评估中，我们在 VE20、VE30、A64FX 和 IceLake-SP 上使用 MPI+OpenMP，在 A100 上使用 MPI+OpenACC。

我们首先使用 SPEChpc 2021基准测试套件中的 微型 工作负载来比较处理器的单插槽性能。最小的微型工作负载由九项基准测试组成，需要约 60GB 内存。我们绘制了 SPEChpc 基准测试脚本报告的每个处理器相对于参考系统（双插槽 12 核 Intel Haswell 系统）的加速比。如果某个处理器由于内存占用要求需要使用多个插槽，则

将加速比除以插槽数量，以进行公平比较。由于撰写本文时 VE30 的编译器仍在开发中，我们无法获得 SOMA 和 Minisweep 在 VE30 上的性能结果。

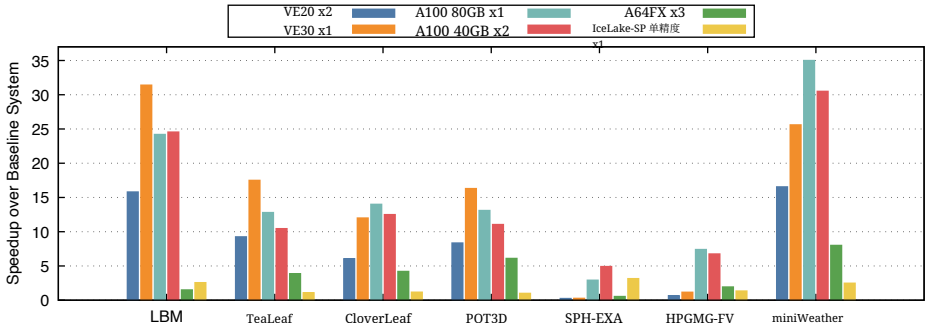


图 11：SPEChpc 2021 微型工作负载性能。

图 11 总结了不同处理器上的 SPEChpc 微型工作负载性能。在 LBM、TeaLeaf 和 POT3D 中，VE30 的性能均优于其他处理器。这三个基准测试相对于 A100 80GB 型号的加速比分别为 1.29×、1.36× 和 1.24×。LBM 和 TeaLeaf 的加速比超过了内存带宽差异所能带来的水平，这表明新引入的 L3缓存以及增加的末级缓存容量和带宽等架构增强正在促进性能提升。

在 CloverLeaf 和 mini- Weather 中，VE30 也明显优于 A64FX 和 IceLake-SP，但其性能略低于 A100 40GB 和 80GB 型号。这是因为 CloverLeaf 中耗时的内核需要执行大量向量聚集操作，与 A100 相比，VE30 似乎难以隐藏向量聚集操作的延迟。miniWeather 基准测试包含内存密集型和计算密集型内核。尽管内存密集型内核在 VE30 上的速度快于 A100，但计算密集型内核在 VE30 上较慢，并且占据了运行时间的主要部分。

SPH-EXA 和 HPGMG-FV 在 VE30 上的表现不佳。SPH-EXA [2] 的主要瓶颈在于构建基于八叉树的空间索引，以及在该索引上执行最近邻查询。这两项功能都需要递归函数调用，无法向量化。要在向量处理器上获得更好的性能，需要将最近邻搜索改为向量友好算法。

HPGMG-FV受循环长度较短的影响。HPGMG-FV微型工作负载将一个 512^3 立方领域分解为 32^3 立方盒子，并将这些盒子分配给MPI进程。耗时最多的Gauss-Seidel红黑平滑器内核遍历一个盒子，其中包含一个三重嵌套循环，每一重循环分别对应一个空间维度。因此，每重循环仅运行32次；但与VE的向量长度相比，这一长度过短，而VE的向量长度为256个双精度元素。一种潜在的优化方法是折叠嵌套循环

并增加循环长度。另一种可能的优化方法是将粗网格层卸载到VH，并在VE上处理细网格层。

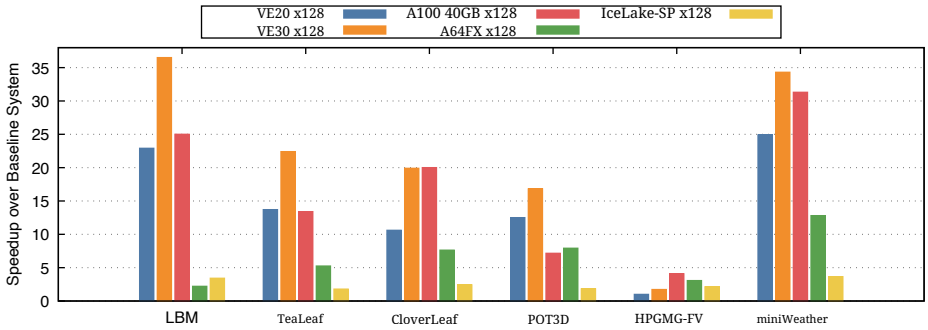


图 12：SPEChpc 2021 中等工作负载性能。

为了评估多节点可扩展性，我们还比较了 SPEChpc中等 工作负载的性能。该中等工作负载由六项基准测试组成，需要约 4TB 内存。在这里，我们在所有处理器上均使用 128 个插槽运行该工作负载。由于我们无法获得 A100 80 GB 型号的大规模部署，因此没有该型号的结果。图 12 总结了中等工作负载的性能。在六项基准测试中，VE30 在其中四项中速度最快，分别是 LBM、TeaLeaf、POT3D 和 miniWeather。与微型工作负载相比，VE30 相对于 A100 的加速比在中等工作负载中更高。例如，在中等工作负载中，VE30 上的 POT3D 比 A100 快 2.34× ，而在微型工作负载中仅快 1.47× 。这表明 VE30 的多节点可扩展性优于 A100。

东北大学内核集合如第 3.3 节所述，东北大学内核集合代表了东北大学网络科学中心用户开发的实际应用。图 13 展示了东北大学内核在 VE20 和 VE30 上的性能。显然，VE30 在所有内核上的表现始终优于 VE20。对于湍流、涡轮机和等离子体，加速比尤为显著，它们在 VE30 上的速度均超过 VE20 的 2.3× 倍。鉴于湍流在 VE20 上受末级缓存带宽限制，我们认为，性能提升得益于末级缓存带宽增加了 2.13× 倍，以及新增的 L3 缓存。如第 3.3 节所述，涡轮机和等离子体则受益于 L3 缓存带来的内存延迟降低。

降雨-径流-淹没模型 降雨-径流-淹没（RRI）模型 [22,26] 是日本广泛采用的二维数值模型，用于进行洪水预报。控制方程采用带有自适应时间步长控制的五阶龙格-库塔方法求解。从计算角度来看，RRI 模型中的主要内核均为内存密集型，因此适合在 VE 上执行。

在本次评估中，我们使用了针对 SX-AT 优化的 RRI 模型实现，并测量了对整个东北 地区日本进行 2 小时洪水预报所需的运行时间。图 14 展示了 VE20 和 VE30 上洪水预报的运行时间。VE30 的性能比 VE20 高 1.32× 。并行区域的加速比为 1.60×。考虑到 VE30 的峰值内存带宽比 VE20 提高了 1.60× ，且 RRI 模型属于内存密集型模型，因此观测到的加速比符合预期。

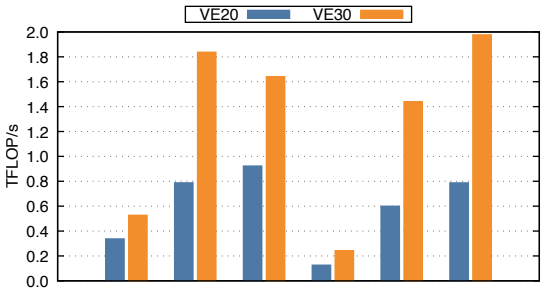


图 13：东北大学内核集合的性能。

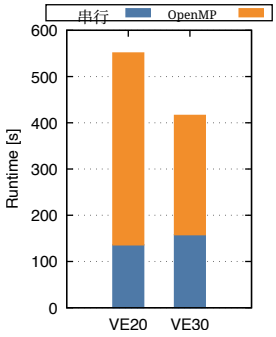


图 14：使用 RRI 模型进行 2 小时洪水预测的运行时间。

4 VE30 性能调优

VE 的基本优化技术包括：将无法向量化的代码从循环中提取出来，以促进向量化；通过各种循环变换增加向量长度；以及将无法向量化的计算卸载到 VH。除了这些优化技术外，还可以利用 VE30 引入的架构特性进一步提升性能。本节将介绍此类调优技术，并量化其对性能的影响。

4.1 选择性L3缓存

在 VE30 上，程序员可以利用可绕过的 L3缓存，有选择地缓存频繁重复使用的数据。为展示选择性 L3 缓存的效果，我们以 Himeno 基准测试为例。代码清单 1.3 展示了 Himeno 基准测试中耗时的 Jacobi 内核。数组 a、b、c、wrk1 和 bnd 按顺序访问，且不会重复使用。另一方面，数组 p 则以类似模板的方式访问。虽然理想情况下，对 p 的 19 次访问中有 18 次应命中缓存，但对其他数组的访问会污染缓存，降低 p 的缓存命中率。只缓存 p，并在访问 a、b、c、wrk1 和 bnd 时绕过缓存，可以缓解这种缓存污染。

列表 1.3：Himeno 基准测试中的雅可比方法内核。

```
用于（i=1 ; i<imax-1;  ++i）
  对于(j=1 ;j<jmax-1 ; ++j)
    for(k=1 ; k<kmax-1 ; ++k){
      s0 = a[0][i][j][k]  * p[i+1][j ][k ] + a[1][i][j][k]   * p[i ][j+1][k ]
         + a[2][i][j][k]  * p[i ][j ][k+1]
         + b[0][i][j][k]  * (p[i+1][j+1][k ]   - p[i+1][j-1][k ]
                          - p[i-1][j+1][k ] +p[i-1][j-1][k ] )
         + b[1][i][j][k]  *  (p[i ][j+1][k+1]  - p[i ][j-1][k+1]
                          - p[i ][j+1][k-1] +p[i ][j-1][k-1] )
         + b[2][i][j][k]  * (p[i+1][j ][k+1]   - p[i-1][j ][k+1]
                          - p[i+1][j ][k-1] +p[i-1][j ][k-1] )
         + c[0][i][j][k]  * p[i-1][j ][k ] +c[1][i][j][k]     * p[i ][j-1][k ]
         + c[2][i][j][k]  * p[i ][j ][k-1] +wrk1[i][j][k];
      ss = (s0   * a[3][i][j][k]  - p[i][j][k] )  * bnd[i][j][k];
      wgosa += ss*ss;
      wrk2[i][j][k] = p[i][j][k] + omega    * ss;
      // 将 wrk2 复制到 wrk，并对所有秩的 wgosa 求和
    }
```

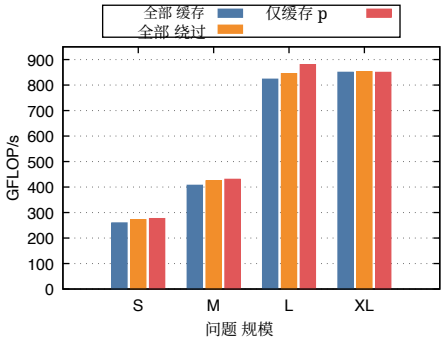
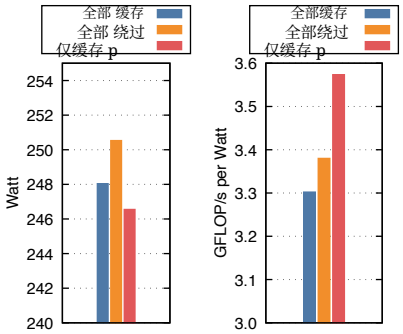


图 15：不同问题规模和 L3 缓存策略下 Himeno 基准测试的性能。



(a) 功耗 (b) 能效

tion

图 16：Himeno 基准测试的能效（L规模）。

图15比较了 Himeno 基准测试在三种不同缓存策略下的性能：(1) 将所有数组缓存到 L3缓存中，(2) 始终绕过 L3缓存，以及 (3) 仅将 p 缓存到 L3缓存中。我们还比较了四种不同的问题规模：S（$2^6 \times 2^6 \times 2^7$）、M（$2^7 \times 2^7 \times 2^8$）、L（$2^8 \times 2^8 \times 2^9$）和 XL（$2^9 \times 2^9 \times 2^{10}$）。结果表明，将所有数组缓存起来与绕过所有数组相比，并未带来显著提升。这说明 L3缓存受到非时间性数据的污染，因此 p 的缓存命中率较低。对于 S、M 和 L 问题规模，与缓存所有数组相比，选择性缓存 p 分别提升了 6.5%、5.7% 和 6.9%。这表明选择性缓存能够缓解缓存污染。相比之下，XL问题规模下未观察到性能提升，这是因为在 XL 规模下，p 无法装入 L3缓存。

为研究选择性缓存是否会影响功耗和能效，我们使用 NEC监控与维护管理器（MMM）[4] 工具，测量运行 Himeno 基准测试时 VE30 PCIe卡的功耗。图 16 比较了不同缓存策略下 VE30 卡的功耗和效率。结果表明，与缓存所有数组相比，仅对 p 进行选择性缓存可使功耗降低 0.6%，这是因为内存访问次数减少了。结合性能提升，选择性缓存使能效提高了 8.2%，达到 3.57 GFLOP/s/W。相比之下，VE20 的能效为 2.21 GFLOP/s/W，A100 40GB型号的能效为 2.14 GFLOP/s/W， [12], VE30 的能效分别高出 1.61× 和 1.66×。

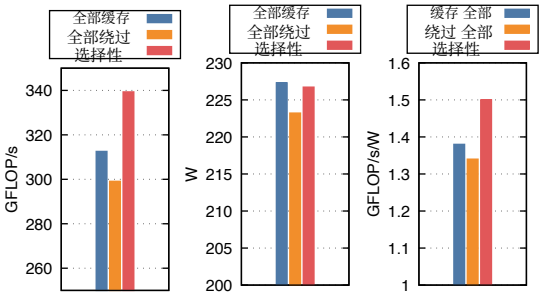
此外，我们将选择性 L3 缓存应用于 Land Mine 内核，以研究选择性缓存是否有利于实际应用。图 17 展示了不同 L3 缓存策略下 Land Mine 内核的性能、功耗和能效。绕过 L3缓存时性能最低，仅为 299GFLOP/s。启用缓存后，性能略微提升至 312GFLOP/s，而选择性缓存进一步将性能提升至 339 GFLOP/s。在功耗方面，缓存所有数组和选择性缓存的功耗都略高于绕过缓存。这是因为缓存功耗的增加超过了内存功耗的降低。不过，选择性缓存带来的性能提升足够显著，因此其能效最高。

4.2 分区模式

如第 2.2 节所述，分区模式通过缓解连接核心与末级缓存的片上网络中的拥塞，提高有效LLC带宽。因此，启用分区模式可能会加速缓存密集型应用。尽管上一代 VE 已支持分区模式，但由于核心数量增加导致片上网络拥塞加剧，预计分区模式在 VE30 上的收益会更大。

为评估分区模式对 VE30 的影响，我们测量了启用和未启用分区模式时 Himeno 基准测试的性能。结果如图

[4] https://sxauroratsubasa.sakura.ne.jp/documents/guide/pdfs/InstallationGuide_E.pdf



(a)性能 (b) 功耗对- （c）能效-
能耗 率

图17：不同L3缓存策略下 Land Mine 内核的性能和能效。

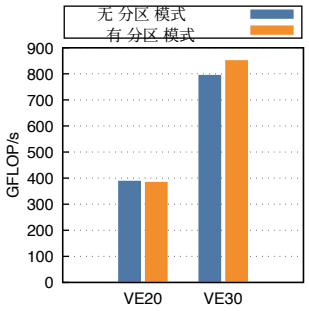


图18：分区模式对 Himeno 基准测试性能的影响（XL 规模）。

18 所示。正如预期，由于片上网络未发生拥塞，分区模式对 VE20 没有显著影响。相比之下，在 VE30 上启用分区模式后，性能提升了 7.1%。这表明，分区模式缓解了片上网络拥塞。因此，在 VE30 上运行缓存密集型应用时，应考虑使用分区模式。

5 结论

本文对搭载全新 VE30 处理器的新一代 SX-AT 超级计算机进行了全面的性能评估。VE30 在 Himeno 基准测试等内存密集型标准基准测试中表现出卓越性能，优于其他处理器。VE30 相较其他处理器的加速比超过了峰值计算性能和内存性能之间的差距，这表明 VE30 引入的新型架构增强带来了显著优势。VE30 在 SPEChpc 等许多实际应用中的表现也优于其他处理器。最后，我们介绍了充分发挥 VE30 潜力的性能调优技术。

这些评估结果清楚地表明，VE30 能够实现可与最新 GPU 和中央处理器媲美的高持续性能，同时允许程序员使用传统编程模型即MPI+OpenMP。这证明，新一代 SX-AT 是寻求实际应用性能的用户的理想选择。

致谢

本研究部分得到了日本学术振兴会科研基金项目编号 JP20H00593、JP20K19808、JP21H03449 和 JP22K19764 的资助。部分实验使用了东北大学网络科学中心的 *AOBA-A*、*AOBA-C*，名古屋大学信息技术中心的 *Flow*，以及大阪大学网络媒体中心的 *SQUID*。

参考文献

1. Ariyoshi, K., Matsuzawa, T., Hasegawa, A.：控制俯冲板块边界上震后滑移传播速度空间变化的关键摩擦参数。Earth and Planetary Science Letters**256**(1-2)，136–146（2007） 2. Cavelan, A., Cabezón, R.M., Grabarczyk, M., Ciorba, F.M.：面向百亿亿次级计算的平滑粒子流体动力学迷你应用。高性能计算、网络、存储与分析国际会议（PASC'20）第 1–11 页（2020 年 6 月） 3. Choquette, J., Gandhi, W., Giroux, O., Stam, N., Krashinsky, R.：NVIDIA A100 张量核心 GPU：性能与创新。IEEE Micro**41**(2)，29–35（2021） 4. Dongarra, J., Heroux, M.A., Luszczek, P.：高性能共轭梯度基准测试：用于高性能计算系统排名的新指标。International Journal of High Performance Computing Applications**30**(1)，3–10（2016） 5. Dongarra, J.J., Luszczek, P., Petite, A.：LINPACK 基准测试：过去、现在与未来。Concurrency a nd Computation: Practice and Experience**15**(9)，803–820（2003） 6. Egawa, R., Fujimoto, S., Yamashita, T., Sasaki, D., Isobe, Y., Shimomura, Y., Takizawa, H.：挖掘第二代 SX-Aurora TSUBASA 的潜力。载于：高性能计算机的性能建模、基准测试与模拟（PMBS 2020）。第 2 卷，第 39–49 页（2020） 7. Himeno, R.：Himeno 基准测试，https://i.riken.jp/en/supercom/documents/ himenobmt/ 8. Hsu, K.C., Tseng, H.W.：使用边缘张量处理单元加速应用。载于：高性能计算、网络、存储与分析国际会议（SC21）。第 1–14 页（2021 年 11 月） 9. Jouppi, N.P., Hyun Yoon, D., Ashcraft, M., Gottscho, M., Jablin, T.B., Kurian, G., Laudon, J., Li, S., Ma, P., Ma, X., Norrie, T., Patil, N., Prasad, S., Young, C., Zhou, Z., Patterson, D.：三代技术塑造 Google TPUv4i 的十条经验。载于：第 48 届年度计算机体系结构国际研讨会（ISCA）。第 1–14 页（2021 年 6 月） 10. Katoh, Y., Ono, T., Iizima, M.：外辐射带中高能电子共振散射的数值模拟。Earth, Planets and Space**57**(2)，117–124（2005） 11. Komatsu, K., Momose, S., Isobe, Y., Watanabe, O., Musa, A., Yokokawa, M., Aoyama, T., Sato, M., Kobayashi, H.：SX-Aurora TSUBASA 矢量超级计算机的性能评估。载于：高性能计算、网络、存储与分析国际会议（SC18）。第 685–696 页（2018 年 11 月） 12. Komatsu, K., Onodera, A., Focht, E., Fujimoto, S., Isobe, Y., Momose, S., Sato, M., Kobayashi, H.：矢量计算系统的性能与功耗分析。Supercomputing Frontiers and Innovations**8**(2)，75–94（2021） 13. Louw, T., Mcintosh-Smith, S.：使用 Graphcore IPU 执行传统高性能计算应用。载于：第 3 届加速机器学习研讨会（AccML）（2021 年 1 月） 14. Mittal, S.：缓存旁路技术综述。Journal of Low Po wer Electronics and Applications **6**(2)（2016） 15. Onodera, A., Komatsu, K., Fujimoto, S., Isobe, Y., Sato, M., Kobayashi, H.：针对 SX-Aurora TSUBASA 的 Himeno 基准测试优化。Lecture Notes in Computer Science（包括 Lecture Notes in Artificial Intelligence 和 Lecture Notes in Bioinformatics 子系列）**12614LNCS**，127–143（2021） 16. Papazian, I.E.：新一代英特尔® 至强® 可扩展处理器。载于：Hot Chips Symposium（2020） 17. RISC-V Foundation：RISC-V "V" 矢量扩展。技术报告（2021），https://github. com/riscv/riscv-v-spec/releases/tag/v1.0

18. Rocki, K., Essendelft, D.V., Sharapov, I., Schreiber, R., Morrison, M., Kibardin, V., Portnoy, A., Dietiker, J.F., Syamlal, M., James, M.：在晶圆级处理器上快速进行模板代码计算。载于：高性能计算、网络、存储与分析国际会议（SC20）（2020年11月） 19. Sato, H., Takagi, Y., Sawaya, K.：具有低交叉极化的高增益对跖费米天线。IEICE通信学报**E94-B**(8), 2292–2297（2011） 20. Sato, M., Kobayashi, T., Zeng, Z., Fang, G., Feng, X.：用于地雷探测的高分辨率探地雷达系统。载于：国际地雷及未爆炸物探测、清除与销毁要求和技术会议论文集。第548–553页（2003） 21. Sato, M., Ishikawa, Y., Tomita, H., Kodama, Y., Odajima, T., Tsuji, M., Yashiro, H., Aoki, M., Shida, N., Miyoshi, I., Hirai, K., Furuya, A., Asato, A., Morita, K., Shimizu, T.：A64FX众核处理器与“Fugaku”的协同设计。载于：高性能计算、网络、存储与分析国际会议（SC20）。第1–15页（2020年11月） 22. Shimomura, Y., Musa, A., Sato, Y., Konja, A., Cui, G., Aoyagi, R., Takahashi, K., Takizawa, H.：基于SX-Aurora TSUBASA的实时洪水淹没预测。载于：第29届高性能计算、数据与分析国际会议（HiPC）（2022） 23. Soga, T., Musa, A., Shimomura, Y., Egawa, R., Itakura, K., Takizawa, H., Okabe, K., Kobayashi, H.：使用实际科学与工程应用评估NEC SX-9的性能。载于：高性能计算、网络、存储与分析国际会议（SC'09）。第1–12页（2009年11月） 24. 标准性能评估公司：SPEChpc 2021（2021），https://www.spec. org/hpc2021/ 25. Stephens, N., Biles, S., Boettcher, M., Eapen, J., Eyole, M., Gabrielli, G., Horsnell, M., Magklis, G., Martinez, A., Premillieu, N., Reid, A., Rico, A., Walker, P.：ARM 可扩展向量扩展。IEEE Micro **37**(2), 26–39（2017） 26. 国际水灾害与风险管理中心：降雨-径流-淹没（RRI）模型，https://www.pwri.go.jp/icharm/research/rri/index.html 27. Tsukahara, T., Iwamoto, K., Kawamura, H.：湍流通道流中的物质线演化。载于：第五届湍流与剪切流现象国际研讨会。第549–554页（2007）