Anton 2：提高专用分子动力学超级计算机的性能和可编程性标准

——国防科大2020年高性能评测与优化课程小组讨论

3-周煜琨　张智超　戴屹钦

指导老师：龚春叶、甘新标、杨博

一、需求分析

- Anton 2是用于分子动力学模拟的第二代专用超级计算机，与其前身Anton 1相比，它在性能、可编程性和容量方面都有显著提高。Anton 2的体系结构是为细粒度事件驱动操作量身定做的，通过增加计算与通信的重叠来提高性能，还允许更广泛的算法高效运行，从而实现许多新的基于软件的优化。目前正在运行的512个节点的Anton 2机器在相同节点数量的情况下比Anton 1快10倍，极大地扩大了全原子生物分子模拟的覆盖范围。Anton 2是第一个为拥有数百万个原子的系统实现每天多微秒物理时间模拟速率的平台。该机器以85μs/天的速度模拟标准的23，558个ATOM基准系统，显示出强大的可扩展性，比任何商用硬件平台或通用超级计算机都快180倍。

二、动机

- 分子动力学(MD)模拟提供了在原子细节水平上对生物大分子行为的特殊可见性。这些模拟的计算需求历来限制了它们的长度，但是并行算法和专用硬件的进步在最近几年已经将这种模拟的覆盖范围扩展到更长的时间尺度。由这一新能力实现的研究--包括几项基于毫秒级全原子MD模拟导致了许多科学发现这些发现涉及以前计算和实验研究都无法获得的生物现象。用于生物分子模拟的化学系统(例如，水中的蛋白质或膜中的蛋白质)的大小从数千个原子到数百万个原子不等，重要生物事件的时间尺度从微秒到几秒的物理时间不等。

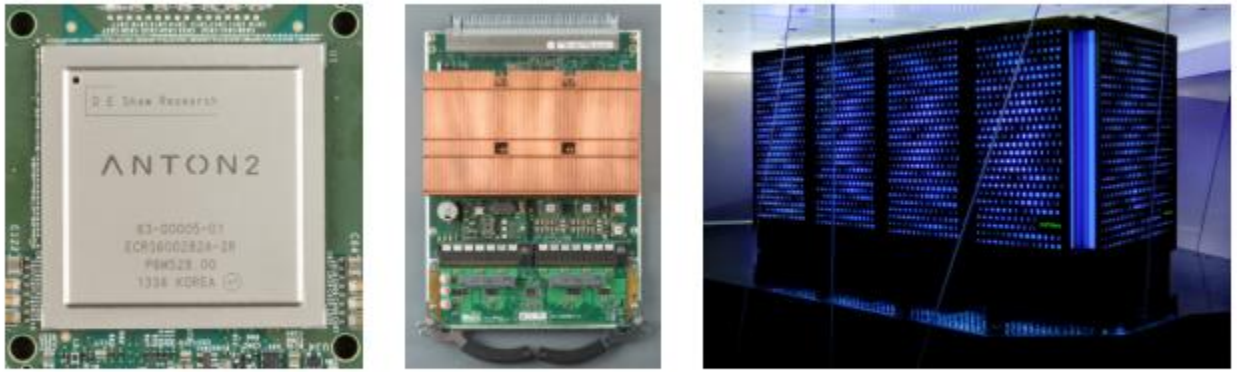
二、动机

- 对于这些系统，即使在最强大的传统超级计算机上，在商用硬件或通用超级计算机上最好的MD软件实现也被限制在每天数百纳秒的物理时间的模拟速率上。使用专用硬件，MD模拟可以加速一个数量级以上：第一代Anton机器(这里称为Anton 1)能够每天模拟10，000到500，000个原子范围内的化学系统。进一步的性能改进将通过极大地减少百微秒范围内的模拟所需的时间，通过对更大、更复杂的化学系统进行毫秒级的模拟，以及通过将更长的时间尺度带到实际范围内，来扩大MD模拟的用途。

二、动机

- 我们设计和建造了Anton 2，这是一台用于MD模拟的专用超级计算机，其性能比Anton 1高出一个数量级，比目前可用的通用硬件高出两个数量级，同时提供了比Anton 1更简单、更灵活的编程模型。与其前身一样，Anton 2在通过专门的高性能网络紧密互连的定制ASIC内执行整个MD计算。每个ASIC将四分之一的芯片面积投入到专门的硬件流水线上，用于计算原子对之间的相互作用，并且还包含66个通用可编程处理器内核，将数据传送到这些流水线并执行MD模拟所需的剩余计算。超大规模集成电路芯片制造技术的改进为Anton 2提供了比Anton 1更多的计算单元，但由于将工作分配给和协调更多单元的管理费用，仅凭这一点不足以提供成比例的更好的性能。因此，Anton 2设计的一个关键组件是一套致力于高效细粒度操作的新机制。由此产生的体系结构更积极地利用了MD模拟的并行性，MD模拟基本上由涉及单个原子或小原子组的大量细粒度计算组成。通过为细粒度通信和同步提供直接硬件支持[20，52]，Anton 2允许这些计算分布在更多数量的功能单元上，同时保持底层硬件资源的高利用率。

二、动机



(a) (b) (c)

Fig. 1 (a) An Anton 2 ASIC. (b) An Anton 2 node board, with one ASIC underneath a phase-change heat sink. (c) A 512-node Anton 2 machine, with four racks.

三、技术方案

TABLE 1. BLOCK-LEVEL COMPARISON OF ANTON 1 AND ANTON 2 ASICS

	Anton 1	Anton 2
Process technology	90 nm	40 nm
Clock speed (GC/PPIM)	485/970 MHz	1.65/1.65 GHz
# of general-purpose processor cores	13	66
# of PPIMs	32	76
Total SRAM + data cache	384 KB	5,280 KB
HTIS memory capacity (atoms)	6,144	32,768
Total data bandwidth to torus neighbors	221 Gb/s	1,075 Gb/s

表1提供了Anton 1和Anton 2 ASIC的块级比较。节点间带宽的提高归功于更积极的信令技术(14 Gb/s对4.6 Gb/s)、更多用于节点间信令的总引脚(384对264)以及更高效的物理层编码。片上SRAM和数据高速缓存的显著增加(14倍)使得为Anton 2编写嵌入式软件变得更容易，同时显著增加了适合片上SRAM的化学系统大小范围。除了这些块级改进之外，以下各节中描述的许多体系结构创新都有助于提高Anton2的性能和可编程性。

三、技术方案

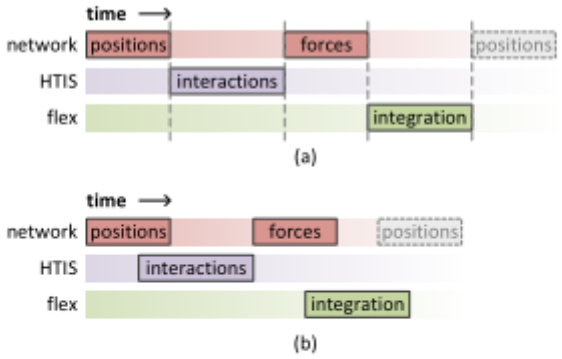


Fig. 5. A range-limited time step with no bond terms consists of (1) sending atom positions to HTIS tiles, (2) computing pairwise interactions, (3) returning forces to flex tiles, then (4) integrating atom positions and velocities, after which positions are sent for the next time step. (a) With a coarse-grained implementation, there is no overlap between communication and computation. (b) A fine-grained implementation allows communication and computation to be substantially overlapped, with the exception that forces cannot be returned from the PPIM array until all interactions have been computed.

图5.没有键合项的范围限制时间步包括(1)将原子位置发送到HTIS块，(2)计算成对相互作用，(3)将力返回力到挠性块，然后(4)对原子位置和速度进行积分，之后将位置发送到下一个时间步。(A)对于粗粒度实现，通信和计算之间没有重叠。(B)细粒度实现允许通信和计算基本上重叠，例外情况是，在计算完所有交互之前，无法从PPIM数组返回力。

三、技术方案

Fig. 6. Event-driven computation on Anton 2. (1) A counted remote write arrives over the network. (2) After the write is processed by SRAM, one or more counters within the dispatch unit are incremented. (3) The geometry cores query the dispatch unit for tasks that are ready to execute (as determined by counters that have reached their thresholds).

图6.Anton 2上的事件驱动计算。(1)计数的远程写入通过网络到达。(2)SRAM处理写入后，递增调度单元内的一个或多个计数器。(3)几何核心向调度单元查询准备好执行的任务(由已达到其阈值的计数器确定)。

三、技术方案

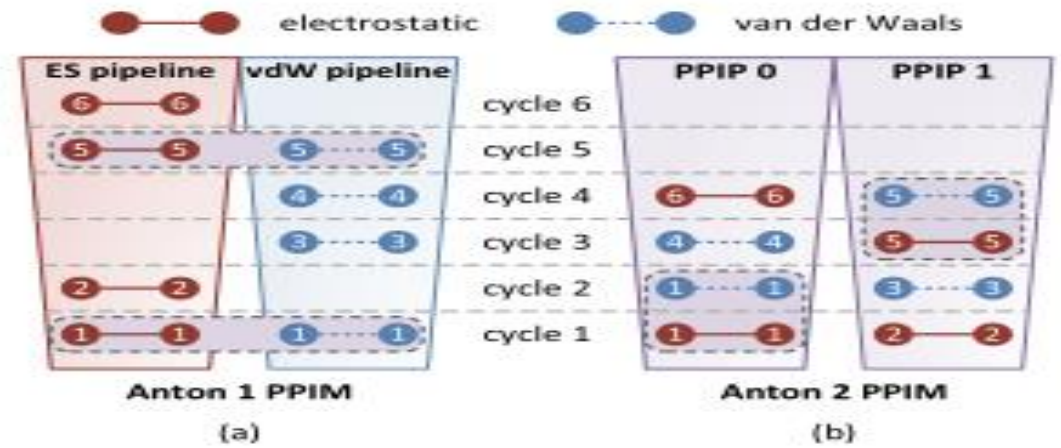


Fig. 7. (a) In the Anton 1 PPIM, electrostatic and van der Waals forces between a pair of atoms are computed on the same cycle by two different pipelines. If one of these forces is not required for a pair of interacting atoms, then the corresponding pipeline is unused during that cycle. (b) The Anton 2 PPIM contains two identical PPIPs. The electrostatic and van der Waals forces between a pair of atoms are computed on consecutive cycles by the same PPIP, improving pipeline utilization.

图7。(A)在Anton 1 PPIM中，一对原子之间的静电力和范德华力是由两个不同的管道在同一周期内计算的。如果一对相互作用的原子不需要这些力中的一种，则在该周期内不使用相应的管道。(B)Anton 2 PPIM包含两个相同的PPIP。一对原子之间的静电和范德华力由相同的PPIP在连续的循环中计算，从而提高了管道利用率。

三、技术方案

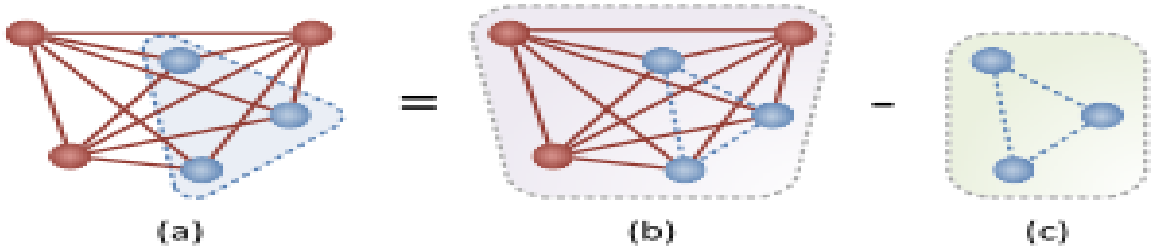


Fig. 8. (a) Three atoms (within the shaded triangle) participate in a bond term. Electrostatic and van der Waals forces (solid lines) are excluded between pairs of these atoms. (b) If the PPIM array computes these undesired interactions (dashed lines), then they must be (c) explicitly recomputed elsewhere and subtracted out. Anton 2 uses topological IDs to avoid computing the majority of these undesired interactions.

also be retrieved from a two-dimensional table instead of being computed from the individual atom parameters, further increasing flexibility by allowing the computed parameters to be overridden on a per-atom-type basis. In particular, this could allow the PPIM to support coarse-grained models such as the MARTINI force field [33].

图8(A)三个原子(在阴影三角形内)参与键项。这些原子对之间不包括静电和范德华力(实线)。(B)如果PPIM数组计算这些不需要的交互(虚线)，则必须(C)在其他地方显式重新计算并减去它们。Anton 2使用拓扑ID来避免计算大多数这些不需要的交互。

三、技术方案

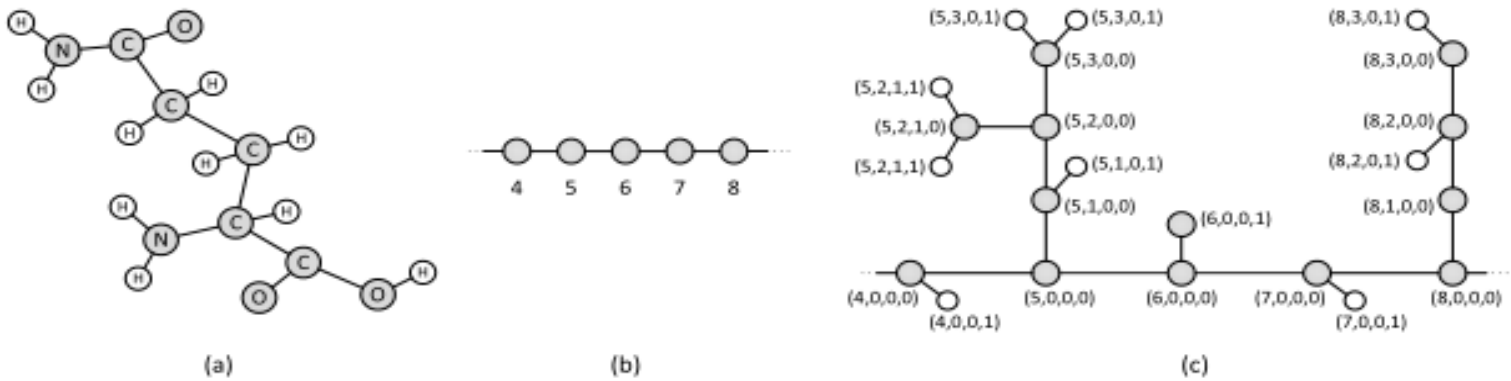


Fig. 9. (a) The bond graph of a molecule (shown for glutamine) is defined by the covalent bonds between atoms. (b) A linear graph could be labeled with consecutive integers; the topological distance between two atoms is just the absolute difference of their labels. (c) Example subgraph labeled with actual topological IDs (n, m, k, t) where n is the backbone index, m is the side-chain index, k is the secondary side-chain index, and $t \in \{0, 1\}$ is a "terminal" flag.

图9(A)分子的键图(如谷氨酰胺所示)是由原子间的共价键定义的。(B)线性图可以用连续整数标号，两个原子之间的拓扑距离就是它们标号的绝对值之差。(C)用实际拓扑ID(n，m，k，t)标记的示例子图，其中n是主干指标，m是侧链指标，k是次要侧链指标，t？{0，1}是“终端”标志。

三、技术方案

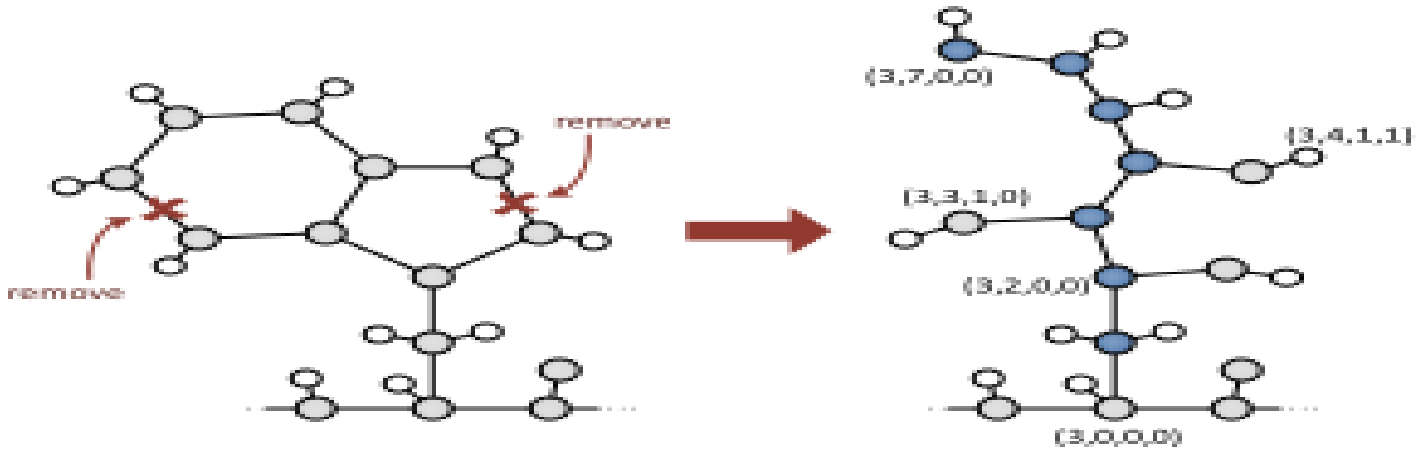


Fig. 10. The bond graph for a tryptophan residue (left) contains cycles, so it must be modified before topological IDs are assigned. Two edges are removed as indicated, resulting in the graph on the right. The side chain within the modified graph is highlighted, and several topological IDs are shown.

图10.色氨酸残基的键合图(左)包含环，因此必须在分配拓扑ID之前对其进行修改。如图所示，删除了两条边，从而产生右侧的图形。修改后的图形中的侧链将高亮显示，并显示几个拓扑ID。

四、效果

- 细粒度操作通过分布式共享内存和事件驱动编程模型向软件公开，并具有用于调度和分派小型计算任务的硬件支持。我们用于Anton 2的MD软件利用了这些通用机制和新算法，这些机制由许多顺序相关的任务组成，这些任务在Anton 1上是不切实际的，但在Anton 2上提供了额外的性能改进。与对细粒度操作的支持相结合，Anton 2包含更专用于加速MD模拟的架构改进。Anton 1和MDGRAPE[34，50，51]都使用专门的硬件管道来显著加快原子对之间相互作用的计算速度。Anton 2采用了相同的方法，通过消除由于计算中的不一致而造成的许多浪费周期，提高了这些管道的利用率。此外，Anton 2使用了一种新的拓扑ID机制来避免计算某些不需要的相互作用。512节点的ANTON 2目前正在运行(图1)，与任何其他现代超级计算机(包括ANTON 1)相比，它在MD模拟方面有三大进步：

四、效果

- 1.峰值性能。在512节点的机器上，ANTON 2实现了二氢叶酸还原酶(Dhfr)的模拟速率为85μs/天，这是一个基准系统，包含23，558个原子，接近实际并行度的极限，因为它每个处理器核心只有不到一个原子(512节点的ANTON 2机器包含33，792个处理器核心)。这一速度比Anton 1提高了4.5倍，比在其他平台上实施MD的速度快180倍。

四、效果

- 2.吞吐量。在较大的化学系统上，ANTON 2的性能比ANTON 1(具有相同节点数)提高了大约10倍，后者对于固定的化学系统规模提供了更高的模拟速率，或者对于原子数量超过10倍的化学系统提供了相同的模拟速率。

四、效果

- 3.1+μs/天的容量。ANTON 2打破了百万原子系统每天微秒的障碍，允许对更大的生物分子，如核糖体进行更长时间的模拟。220万个原子的核糖体模拟在512个节点上以3.6Gbps/天的速度运行，比在SuperMUC群集的1，024个节点上模拟类似的核糖体系统快21倍[28]。在更大的Anton2机器上可以获得更高的性能：虽然迄今为止构建的最大机器包含512个节点，但该体系结构可扩展到4096个节点。

五、分析

- ANTON 2提高了专用分子动力学超级计算机的性能和可编程性的技术水平。对于各种系统大小，在Anton 2上的模拟运行速度比在Anton 1上快4-10倍，比在任何通用硬件上快两个数量级以上。实现这一性能需要多种技术，包括针对细粒度操作进行优化的体系结构、更高效的硬件流水线、新颖的拓扑ID机制，以及一组由硬件支持的事件驱动计算算法。在Anton 2上性能提升的同时，灵活性也得到了提高。

五、分析

- 除了使用最常用的MD力场加速模拟之外，Anton 2还包含对各种力场扩展的直接支持，包括粗粒度模型和非标准成对交互。这种支持，再加上改进的可编程性，使Anton 2成为交替力场和力场研究的强大平台。虽然Anton 2旨在加速MD模拟，但它采用的许多机制-即支持高效的细粒度通信、计算和同步-将有利于更广泛的科学计算类别。

五、分析

- 举三个例子，像神经网络、稀疏线性代数和流体动力学等各种各样的应用程序都可以自然地表示为大量小型的、事件驱动的计算。使用单个指令发送计数的远程写入、在内存中执行累加以及对单个四元组数据进行同步的能力使编写直接捕获底层计算数据流的软件变得更容易。此外，分派单元允许将计算表示为细粒度任务的集合，而不必在软件中显式地调度这些任务。我们在为Anton 2实施MD软件方面的定性经验是，它的编程比Anton 1容易得多，这主要是因为它支持细粒度操作，并且使用了一组同构的嵌入式处理器内核。虽然高度专业化，但ANTON 2提供了显著的可编程性和灵活性，这将促进其在推进生物分子模拟的前沿领域的使用。

- 谢谢，欢迎批评指正