白皮书

第四代英特尔® 至强® 可扩展处理器
英特尔® AMX / 英特尔® SGX
英特尔® 傲腾™ 持久内存 300 系列
云服务器/裸金属

第四代英特尔® 至强® 可扩展处理器全新升级，助力百度智能云打造新一代云智一体架构产品

百度智能云

目录

前言概述 1

方案背景：借助云智一体架构，为行业用户提供智能时代基础设施 2

解决方案：百度与英特尔携手，为新一代云智一体产品提供算力与 AI 加持 3

- 全新处理器架构及性能升级，为 AI 应用提供充沛算力 3
- 全新处理器内置 AI 能力，为 AI 应用注入有效加速 4
- 借力傲腾™ 持久内存，满足 AI 计算所需大容量内存 5
- 以英特尔® SGX 为数据安全性提供硬件级保障 5

优化方案实践落地 6

- 应用场景 1：生命科学 6
- 应用场景 2：自动驾驶 6
- 应用场景 3：工业制造 6

未来展望 7

前言概述

随着云服务和人工智能（Artificial Intelligence，AI）在各行各业的核心业务中承担越来越重要的职责，以云计算为基础、以 AI 为引擎，打造云智一体的智能时代基础设施，正逐渐成为各大云服务平台助力行业用户加速智能化升级和转型，赢得市场竞争的重要方法。

作为全球领先的云服务提供商之一，百度智能云也依托百度在 AI 领域的技术优势和经验积累，通过与行业应用的深度融合，带动和沉淀其 AI PaaS（Platform as a Service，平台即服务）层和 AI IaaS（Infrastructure as a Service，基础设施即服务）层能力，以持续的产品优化和升级来为行业用户打造强劲且高性价比的算力支持，以及高效便捷的 AI 开发运行能力。

这一过程中，百度智能云与英特尔开展了一系列深度技术合作，并在最新一代的云智一体架构产品，第六代云服务器（BCC）、裸金属（BBC）等产品中引入全新第四代英特尔® 至强® 可扩展处理器作为各项“云 + AI”需求的算力引擎。除基础算力得到大幅提升之外，新处理器内置的英特尔® 高级矩阵扩展（Intel® Advanced Matrix Extensions，英特尔® AMX）也能帮助实例获得有效的 AI 加速。除此之外，英特尔® 傲腾™ 持久内存 300 系列、英特尔® 软件防护扩展（英特尔® Software Guard Extensions，英特尔® SGX）等产品与技术的引入，也将帮助百度智能云更好地应对用户场景中对大容量内存、数据安全等的需求与挑战。

> 未来的云服务将在实现各行各业数字化转型的基础上，聚焦并持续深化推动其智能化升级进程。我们基于云智一体 3.0 架构的第六代云服务器、裸金属等产品，就是从 AI IaaS 层和 AI PaaS 层能力入手，提供用户所需的极致算力和高效 AI 开发能力。第四代英特尔® 至强® 可扩展处理器、英特尔® AMX 等产品与技术，为我们的新产品提供了从算力、AI 加速、内存扩容到数据安全等的全面加持。
>
> **谢广军**
> **百度副总裁**
> **百度智能云**

方案背景：借助云智一体架构，为行业用户提供智能时代基础设施

得益于云计算技术的高速发展，各类云平台已逐渐成为千行百业开展数字化转型和智能化升级的基座。而原生化 AI 技术的加持，更让云服务对用户核心业务的支持变得高效。例如在制造、金融等传统行业中，AI 可助力用户基于云平台重构从研发设计到业务生产的全流程。而在自动驾驶、生物医疗等新兴产业中，由云平台提供的 AI 能力也成为新场景、新模式构建时的技术基座。

作为全球为数不多，能提供从 AI 芯片、软件架构到应用程序的全栈 AI 技术公司，百度在这一浪潮中，借助其在自研技术、产品和生态上的领先优势，不仅通过百度智能云将“适合跑 AI 的云”提供给用户，助其实现安全、稳定、灵活的数字化转型，也以“懂场景的 AI”这一智能化引擎，为用户的智能化升级和转型提供领先的创新技术和平台。

为了与用户需求形成更深层次的融合，百度智能云计划在其最新的 3.0 版本中带动和沉淀更多 AI PaaS 层和 AI IaaS 层的能力，打造具有性价比的异构算力和高效的 AI 开发运行能力，从而形成有效的智能化闭环路径，包括：

- **AI IaaS:** 基于百舸 AI 异构计算平台 2.0，以异构计算实例、海量数据湖存储与高性能存储等来推动 AI 计算、存储、加速框架及容器等在各业务场景下的落地与升级；

图一 基于云智一体架构的百度智能云

- **AI PaaS:** 基于飞桨深度学习平台、文心行业大模型等提供的强力支撑，助力落地场景充分发挥算力引擎作用，加速模型迭代。

这一过程中，百度智能云也面临着新的需求与挑战，包括：

- **更强劲且支持 AI 加速的算力需求:** AI 应用对算力需求的持续提升，以及对性能功耗比的更多关注，正推动百度智能云在绿色 IDC 建设、功耗优化等措施之外，寻求能支撑多种算力，覆盖更广维度计算场景，且能够有效实现 AI 加速的多元化算力设备，让云计算实例能从“All-in-One”向着“One (XPU)-for-All”转变；
- **更多面向AI计算的内存需求:** AI 模型参数量的持续增大，对内存容量提出更高要求，OOM (Out Of Memory) 问题逐渐显现，内存墙无形中成为一部分 AI 应用的瓶颈。同时，传统 AI 算力设备内存扩容能力有限，且传输速率升级无法兼顾低成本与算力的高速增长，因此百度智能云亟需寻求一种高效、高性价比的方式来予以应对；
- **更高的安全和数据合规性要求:** 元宇宙等融合虚拟资产、混合现实社交元素的新场景的出现，以及企业对数据收集、脱敏、标注等流程安全性的关注，推动百度智能云在数据安全性上投入更多关注，这不仅涉及到算力部署形态的调整，也对算力设备提出了物理级别的安全防护要求。

为此，百度智能云与英特尔携手开展一系列深度技术合作，引入第四代英特尔® 至强® 可扩展处理器、英特尔® AMX、英特尔® 傲腾™ 持久内存 300 系列、以及英特尔® SGX 等产品与技术，打造第六代 BCC / BBC 等产品。通过基础性能、实例丰富度、实例特性等维度的有效提升，为上述需求与挑战提供有效解决之道，使百度智能云能在高品质的算力、AI 加速、内存和数据安全性基础上，更聚焦于对用户丰富应用场景的深耕和对行业关键业务能力的解析，助力用户持续降本增效、在实现业务转型与升级之余也加速其智能化升级的进程。

解决方案：百度与英特尔携手，为新一代云智一体产品提供算力与 AI 加持

云智一体的架构，能让各行业用户在更多“云 + AI”的场景下，借助百度智能云来解决多元化的业务需求，但这也对云平台的效能提出了不同维度的挑战。为此，第六代 BCC / BBC 产品借助各类软硬件产品的优化升级，使实例在跨代同规格上实现了 50% 的综合性能提升，最高性能提升可达 70%[1]。而在此之外，针对用户在算力、AI 加速、内存和数据安全性上的需求，第六代 BCC / BBC 产品也围绕全新的第四代英特尔® 至强® 可扩展处理器这一核心，给出了卓有成效的产品方案。

■ 全新处理器架构及性能升级，为 AI 应用提供充沛算力

在用户所聚焦的算力需求上，第六代 BCC / BBC 产品引入第四代英特尔® 至强® 可扩展处理器来满足用户在不同场景下的需求。作为至强® 可扩展处理器家族的“新秀”，这一新处理器采用Intel 7 制程工艺，可凭借全新的性能核微架构设计来提升处理速度，在低时延和单线程性能上实现突破。在芯片架构层面，新处理器借助嵌入式多芯片互连桥接 (Embedded Multi-die Interconnect Bridge，EMIB) 技术，帮助第六代 BCC / BBC产品在保持既有单核优势的同时也能大幅提升可扩展性。同时，新处理器也提供了对先进内存和下一代 I/O 技术，包括 DDR5、PCIe 5.0、CXL (Compute Express Link) 1.1 以及高带宽内存 (High Bandwidth Memory，HBM) 等技术的支持。

在业内通用应用所关注的 benchmark 项目中，百度智能云第六代实例基于英特尔新一代处理器，表现出了优异的性能升级。

在如 SPEC2017 Rate、Linpack 等通用测试中，新一代实例整体性能相较于上一代产品最高提升达 70%[2]；内存性能上，第六代百度智能云实例相对于上一代产品可实现至少 40% 的内存速率提升[3]，使用户在不调整业务架构的条件下，可更高效地使用内存存储空间。同时，新的互联协议的引入令通信速率提高到 16GT/s，其直观表现即为在本地 SSD 型实例 I6 上，相较于上一代同规格实例实现了约 40% 的 IOPS 提升与至少 15% 的读写时延下降，用户可轻松获得百万级别 IOPS 的存储能力[4]。基于一系列优化，百度智能云将更有效地帮助用户进一步提升其 I/O 密集型业务运行效率。

在高性能计算类应用中，如大规模数据处理、机器学习、科学计算，以及大型前端 Web 服务器等场景中，用户有着更高的数据实时计算处理、更大的网络吞吐率等需求。为此，第六代 BCC/BBC 产品借助第四代英特尔® 至强® 可扩展处理器提供的更多核心数量，为用户提供了 192 至 208vCPU（virtual CPU，虚拟处理器核心）的多核心实例规格族，同时也具备了 2 x 100Gbps 双向互联带宽 VPC（Virtual Private Cloud，私有网络）网络以及超过 3,000 万 PPS（Packets Per Second，每秒数据包）的高网络吞吐率，同时 VPC 网络时延最高可降低 40%，以更强的网络能力匹配算力升级，实现业务性能的整体提升[5]。

在元宇宙、游戏娱乐、渲染平台等计算密集型任务中，对任务线程级的处理速率有着较高要求，这要求算力设备具备更高主频和更高睿频。搭载第四代英特尔® 至强® 可扩展处理器的第六代 BCC / BBC 产品可支持最大 3.1GHz 基频，3.4GHz 全核睿频的高主频实例规格族，让用户能够游刃有余地选择更贴近业务能力的算力底座。

■ 全新处理器内置 AI 能力，为 AI 应用注入有效加速

与此同时，第四代英特尔® 至强® 可扩展处理器内置的英特尔® AMX，也能帮助百度智能云大幅升级 AI 性能。借助英特尔® AMX 全新的指令集与电路设计，其提供的加速能力可以使各类 AI 负载，例如图像识别、对象探测等任务中的张量处理效率获得大幅提升。同时，这一技术可作用在 INT8、BF16 等不

图二 第四代英特尔® 至强® 可扩展处理器

同数据格式上，这使得性能加速能在更多数据格式的 AI 模型上起效，从而让更多使用标准规格族的用户享受到 AI 加速红利，并让 AI 算力拓展到更多实例中。

值得一提的是，为更有效地实现对英特尔® AMX 指令的调用，百度智能云也引入了英特尔® oneAPI 工具套件。作为开源、跨平台的性能库，英特尔® oneAPI 工具套件可有效助力云平台上的用户提升其 AI 应用与框架的性能，并在面向渲染型、高性能计算的实例中提升整体 AI 性能表现。

■ 借力傲腾™ 持久内存，满足 AI 计算所需大容量内存

在目前广泛应用的各个 AI 推理场景，例如金融行业的实时反欺诈场景中，长年累月的数据量增长以及索引规模扩大对 AI 任务中的内存容量有了更大要求。又如在生命科学行业，特别是在大分子类药物合成、分子动力学等场景中，AI 推理对大容量内存需求的急速增长，使百度智能云面临越来越多关于内存容量不足的用户反馈。

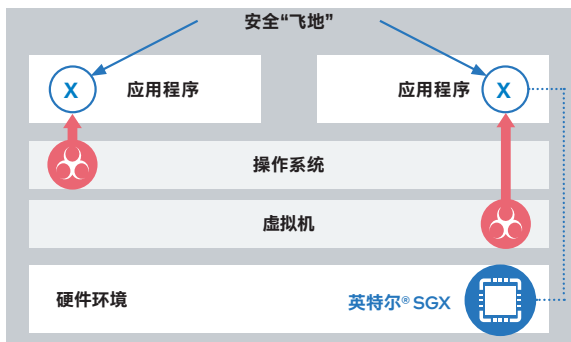
因此在算力提升之外，第六代 BCC / BBC 产品也针对 AI 计算中关键的大容量内存需求，率先引入英特尔® 傲腾™ 持久内存 300 系列。基于这一傲腾™ 持久内存家族的最新产品相较于上一代实例可实现 20% 的最大容量拓展[6]，并可通过对工具库的支持，进一步优化实例的内存成本，推动用户应用以 In-Memory 模式部署，从而获得更优的性能。得益于第四代英特尔® 至强® 可扩展处理器对傲腾™ 持久内存 300 系列的升级支持，这两者的强强组合能帮助第六代 BCC / BBC 产品中大内存型实例的内存总容量达到常规纯 DRAM 内存部署模式的 2 倍以上，同时还能基于其数据持久性的特点提供快速应用恢复的能力[7]。

事实上，在大容量与成本优势之外，英特尔® 傲腾™ 持久内存 300 系列对 CXL 协议的支持和推动作用，也能帮助百度智能云从容应对未来 AI 应用铺开而带来的算力及内存需求指数级增长。作为由英特尔主导的新一代开放性互联协议，CXL 协议能让通用处理器、FPGA（Field Programmable Gate Array，可编程阵列逻辑）等硬件算力设备和加速器实现高速、高效的互联，满足百度智能云用户对高性能异构计算的要求，并提供更高的带宽及更好的内存一致性。

■ 以英特尔® SGX 为数据安全性提供硬件级保障

数据安全正成为政策核心关注点，随着更多 AI 应用，如元宇宙、自动驾驶等在运行过程中对个人或组织敏感信息的更频繁调用，保证数据在收集、脱敏、标注时的安全性也受到了百度智能云的高度重视。

内置于第四代英特尔® 至强® 可扩展处理器中的英特尔® SGX，能在内存等硬件环境中构造出一个可信的安全“飞地”（Enclave），为敏感数据和代码提供独立于操作系统和硬件配置的、增强的安全防护。通过英特尔® SGX 的加持，第六代 BCC / BBC 产品的实例可在很大概率上保证云上运行业务的代码、数据不被 OS、虚机监控器等监视、修改，从而能提供对业务过程数据安全性的重要保障。



图三 英特尔® SGX 实际作用示意图

优化方案实践落地

随着第四代英特尔® 至强® 可扩展处理器等产品与技术在百度智能云第六代 BCC / BBC 产品中获得充分地融合，由第六代 BCC / BBC 产品提供的云实例也在生命科学、自动驾驶以及工业制造等一系列 AI 应用领域获得落地部署，并在实践应用中获得了用户的良好反馈。

应用场景 1: 生命科学

近年来，生命科学领域已成为 AI 技术落地运用的热点行业之一。无论是制药（如智能药物、细胞药）、化工与农业&食品（如生物材料、AI 育种、合成蛋白），还是医疗（如精准治疗方向、脑机）等，都对基于云平台的算力与 AI 加速能力提出了巨大的需求。生命科学领域的“云 + AI”需求有着以下特点：

- 数据安全性要求高，需从计算、存储、网络等多方面提供面向数据流通、数据隐私保护的实现方案；
- 算力需求高，部分场景下需要大规模 AI 集群，因此对集群的整体性能存在要求；
- 对工具组件、端到端开发套件以及基础模型库等方面也有着相对多样性的需求。

百度智能云产品方案： 通过第六代 BCC + BBC 的实例组合，百度智能云以第四代英特尔® 至强® 可扩展处理器为核心，为用户提供多种端到端的生命科学 AI 实例搭配方案。用户可通过搭载英特尔® SGX 的裸金属安全型实例，配合百度智能云对象存储、网络产品来实现全链条数据流通加密。算力方面，可以使用百度智能云高性能计算集群 EHC，以包括第四代英特尔® 至强® 可扩展处理器在内的异构算力资源为基座，配合资源、作业调度以及 AI 加速工具集应对大规模生科算力需求。同时，百度智能云也在高速推动基于英特尔® 架构 SoC（System on Chip，系统级芯片）方案的百度太行智能网卡所驱动的第六代 RDMA 计算增强型实例落地，帮助用户实现高网络能力算力的灵活调度。

应用场景 2: 自动驾驶

自动驾驶场景对云服务能力有着较为严苛的要求，需要提供丰富的实例规格及便捷的部署形态，同时需在安全合规的前提下满足其在数据采集、标注、训练以及仿真等一系列流程中的要求。包括：

- 决策训练过程：处理器的计算效能及实例的 IOPS 是算法模型运行时的核心关注点；
- 仿真过程：更注重处理器等算力设备的渲染和推理等能力。

百度智能云产品方案： 得益于第四代英特尔® 至强® 可扩展处理器优异的表现，第六代 BCC + BBC 提供的实例在单核表现以及核心数量上都有着大幅提升。其中通用型 g 系列（融合了英特尔® AMX）、本地 SSD 型 I 系列都可成为用户在自动驾驶场景下的不二之选，同时实例对 AMX 指令集的支持也提供给用户更多实例选择，通过强大的矩阵运算加速能力加持，一部分业务场景的通用、异构算力可完全基于以上实例顺利驱动。而产品所具备的 2x100Gbps 双向互联带宽 VPC，超过 3,000 万 PPS 的高网络吞吐率，以及理论可达 100GB / s 的超高性能本地存储能力，都将在自动驾驶领域中助力用户实现业务模型的高质量落地[8]。

应用场景 3: 工业制造

在工业制造场景中，不同的业务对云服务也有着不同的需求。例如：

- EFEA（能量有限元分析法）相关的高速冲击分析等问题，通常需要高 I/O 及配套的高网络性能，对实例规格要求较高；
- IFEA（隐式有限元分析）相关的低速冲击分析等问题，对计算中的内存有着较大容量需求；
- CFD（计算流体力学）相关业务对 vCPU 数量要求较高，需满足多线程数需求。

因此，工业制造场景对于云服务实例的全局能力有较高要求，需在各个方向上均有着相对领先的能力以提供足够的场景支撑。

百度智能云产品方案：借助第六代 BCC + BBC 产品提供的实例组合，百度智能云能以丰富的实例规格族完整覆盖工业制造场景下的各类性能要求。相较于上一代实例，第六代百度智能云 BCC / BBC 实例可提供具有 1.6 倍核心数提升、2 倍 I/O 性能飞跃的新实例规格族，完美应对多线程数场景、高 I/O 场景的极致性能要求[9]。同时，针对部分场景对内存的刚性需求，第六代 BBC 产品也将推出搭载英特尔® 傲腾™ 持久内存的内存增强型实例，帮助用户将更多数据部署到内存进行快速计算处理，大幅加速工业制造场景中的数据处理效率。

未来展望

面向未来，百度智能云将继续与英特尔携手，基于 AI 与云服务的技术潮流，推动行业的智能化升级。在长期协作共进的道路上，百度智能云一方面希望继续保持对英特尔® 至强® 可扩展处理器产品在 Built-in AI 演进上的跟踪，并协同英特尔推出更多的实例系列或最佳实践。如本案例中，第四代英特尔® 至强® 可扩展处理器所加入的英特尔® AMX 就能很大程度上帮助百度智能云的用户在通用的实例规格族中高效使用到 AI 推理和甚至部分训练有关的能力，在大幅提高能效比的同时也节约了成本。

另一方面，百度智能云也希望通过引入更多英特尔软硬件产品与技术，帮助用户在专门领域中深挖实例潜力，实现业务增效，并在不同实例中帮助用户更好地利用英特尔的 AI 全环节能力。例如通过英特尔® FPGA 产品来构建场景化 AI 能力，实现降本增效；通过引入英特尔® oneAPI 工具套件，在渲染、高性能计算等场景中提升加速器中工作负载的开发、部署效率，以及通过特定编程模型来实现软、硬件的高效协同等。

一般提示和法律声明

[1, 2, 3, 4, 5, 6, 7, 8, 9] 测试配置：处理器：双路第四代英特尔® 至强® 可扩展处理器；内存：DDR5-4800 64G RDIMM *16；操作系统版本：Centos 7（内核：5.10 版本）。
如欲了解更多详情，请联系百度智能云：https://cloud.baidu.com/。
英特尔并不控制或审计第三方数据。请您审查该内容，咨询其他来源，并确认提及数据是否准确。

关于英特尔的性能和基准测试程序结果的更多信息，请访问 www.intel.com/benchmarks。
英特尔并不控制或审计他人数据。请您审查该内容，咨询其他来源，并确认提及数据是否准确。
英特尔技术特性和优势取决于系统配置，并可能需要支持的硬件、软件或服务得以激活。产品性能会基于系统配置有所变化。没有任何产品或组件是绝对安全的。更多信息请从原始设备制造商或零售商处获得，或请见 intel.com。
英特尔未做出任何明示和默示的保证，包括但不限于，关于适销性、适合特定目的及不侵权的默示保证，以及在履约过程、交易过程或贸易惯例中引起的任何保证。