南開大學
Nankai University

计算机学院

并行程序设计报告

并行体系结构调研

马浩祎

学号：2213559

专业：计算机科学与技术

2024 年 3 月 16 日

目录

1 引言——关于并行体系结构

1.1 概念

并行体系结构是建立在计算机体系结构上的，其在计算机系统上是具有并行性的，即可以同时进行多于两个运算或操作。并行体系结构一般从时间和空间两方面考虑：

- 时间层次：指时间重叠，让多个处理过程在时间上相互错开，以加快硬件运行效率。
- 空间层次：指资源重复，通过超大规模集成电路技术在多处理器系统和多计算机系统中广泛应用。

1.2 发展历史简述

并行计算的历史可以追溯到 1960 年代，当时的大型计算机主要是采用了多个处理单元的设计，也就是多核。随着计算机技术的发展，并行计算在 1970 年代和 1980 年代得到了广泛应用，尤其是在科学计算和工程计算领域。1990 年代以来，并行计算技术的发展加速，并且越来越多的商业应用开始使用并行计算。这期间并行计算也与时俱进，表现出更多且更新的特点，后文会详细阐述。

1.3 具体架构设计

- 分布式系统：多个计算机通过网络相互连接，形成一个大型并行计算系统。
- 共享内存系统：多个处理单元共享同一块内存，可以直接访问彼此的数据。
- 异构系统：多个处理单元采用不同的处理方式，如 GPU、TPU 等。

设计中的技术点：

- 处理单元的组织：如何将多个处理单元组织成一个整体。
- 内存组织：如何将内存分配给多个处理单元。
- 通信方式：如何实现多个处理单元之间的数据交换和同步。
- 负载均衡：如何确保多个处理单元的工作负载相等。

2 超级计算机中的并行

2.1 超级计算机——富岳

2.1.1 基于 ARM 架构的超算集群

如图 2.1，富岳超级计算机系统有 396 个满配的 Rack 和 36 个半配的 Rack，一个 Rack 有 384 个 Node(CPU)，那么 Node 总数就是 158976 个 CPU，由此可看出大量的内核集群是超级计算机高性能计算的基础。

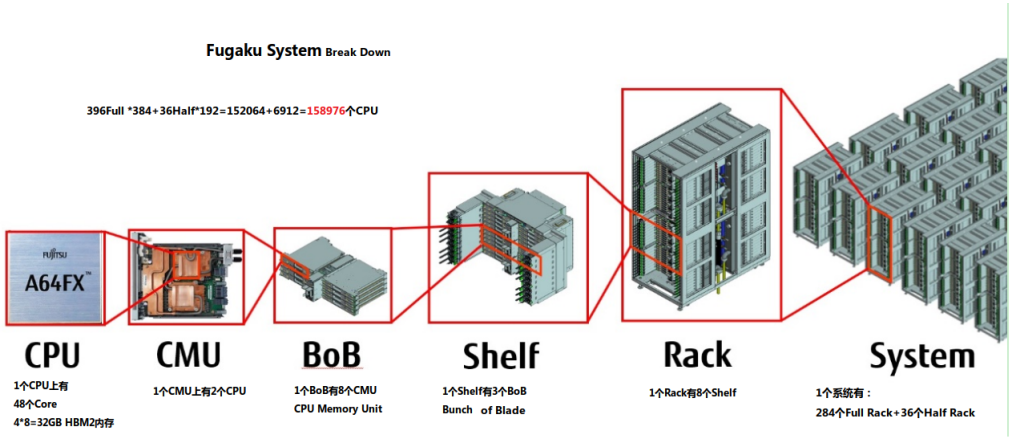

图 2.1: 富岳超级计算机系统架构

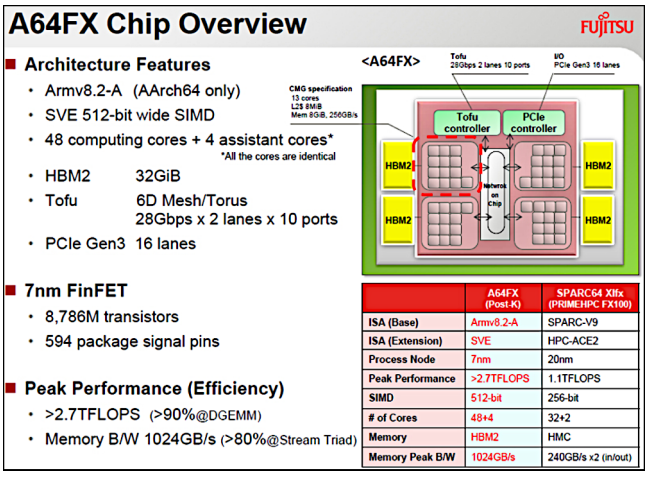

图 2.2: A64FX 处理器

2.1.2 核心处理器：A64FX

如图 2.2，A64FX 是日本富士通发布的“最强”ARM 处理器，在一个 CPU 内部集成了 52 个核心，配备 32GB HBM 2 内存，带宽 1TB/s，浮点性能 2.7TFLOPS，使用 7nm 工艺生产 [1]。我们也可以在图中看出其优越的性能对比。

该 CPU 采用 SVE 可伸缩矢量扩展指令集，支持 512bit 浮点运算单元，浮点计算性能大幅增强，该指令集有如下优点：

- 可伸缩的向量长度：通过选择合适的长度来提高并行度
- 单通道 predication：允许包含复杂控制流的循环矢量化
- Predicate-driven 的循环控制和管理：降低了将标量代码的矢量化开销
- 丰富的并行化操作：适用于更多类型的可分解 loop-carried 依赖关系
- 向量分区和软件管理的推测：支持将数据依赖的 loop 矢量化
- 可伸缩的向量内子循环允许对具有更复杂 loop-carried 依赖关系的循环进行矢量化

2.1.3 互联 TofuD 网络

该互联方案是指在 ARM 架构的 CPU 的基础上进一步定制，使每个 CPU 与一块本地的 HBM 高带宽内存直连，不同的 CPU 通过片上互连构成 NUMA 架构 [2]。同时，芯片上集成了 6 个网络接口，CPU 通过片上互连访问这些网络接口，避免了 PCIe 及 I/O 子系统带来的开销，如图 2.3。同时，每个网络接口和一个路由器连接，外部接入一个 6D Torus 网络。如此做可让延迟下降 50% 左右，具体体现在：

- 网卡被直接集成在芯片内部，取代 Host Bus
- 网卡直接向 CPU Cache 写数据，减少访存开销
- SerDes 带宽的提升可能会造成延迟的增加，故降低带宽以减少延迟

而且该方案通过 6 个 Link 直接大幅提升总带宽，理论可达到 40.8GB/s，实际是 38.1GB/s，转化效率高达 93%。

2.2 现代并行技术

根据上节针对富岳超级计算机的调研分析，结合有关资料 [3] 不难得出现代并行技术的发展趋势。

2.2.1 可扩展性 (Scalability)

我们刚才分析了富岳超算的超大规模处理单元集群，但是可扩展性的含义不只是系统中处理机数目的多少，也不仅是系统可达到接近线性的加速比。实现可扩展性不仅是一种技术，更是一种设计原则，它与系统及应用的规模、性能、成本都有联系，而且与技术更新及产品升级换代有关，涉及资源、应用与技术等多个方面。

可扩展并行机的一个重要发展趋势是从器件级集成过渡到微机、工作站主板甚至整机集成。这种新的构造并行机的方式可以保证并行机与微机或工作站完全同步升级，几乎没有延迟。而且可扩展的系统一定是一个各方面平衡的系统，没有明显的性能瓶颈。我认为，目前该方向有以下关键技术：

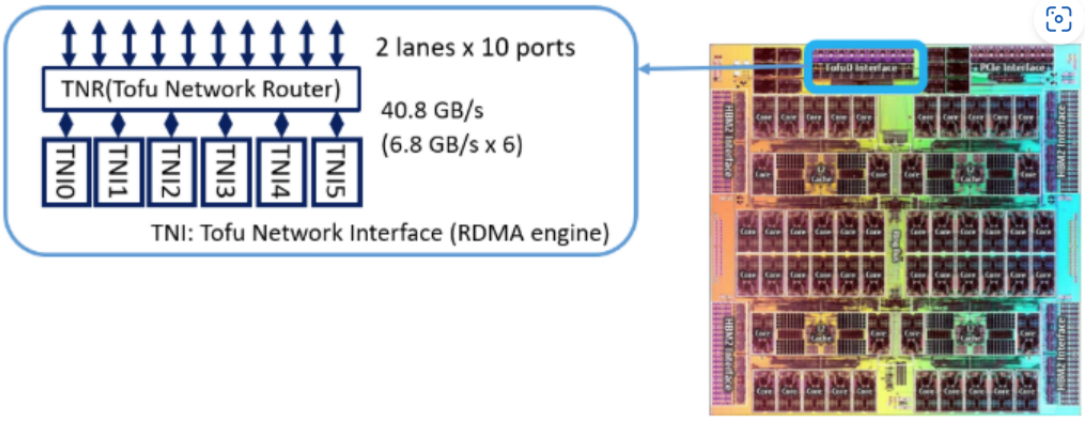

图 2.3: CPU 上的 TofuD 通信网络

- 具有单一系统形象，即统一地址空间，单一文件系统，单一系统形象可以在系统级也可以在用户级实现
- 隐藏延迟是实现可扩展性的一大途径，主要技术是数据预取与多线程机制。
- 编译技术将在可扩展并行机研制中发挥越来越重要的作用。
- 采用多线程机制可实现计算与通信重迭，当一个通信线程挂起等待时，可启动另一计算线程。

总之，可扩展技术是高性能计算机的核心技术，从共享存储多处理机，包含几十个处理机的机群发展到由几百、几千甚至更多处理机组成的可扩展高性能计算机，追求的目标就是系统具有可扩展性。

2.2.2 低延迟高带宽的互连网络接口

上面我们也分析到了超算的通信网络方案，在设计并行架构时，当我们有了很多处理单元后，怎么高效做到单元之间的连接也是一个关键性的技术问题，这里我关注到以下几点：

- 硬件延迟其实不是很大，不到 1 微秒，但是通信软件因为价格昂贵，使得用户级的通信延迟在高达毫秒级，致使效率过低。
- 传统网络接口采用的 DMA 机制，需要经过操作系统，导致系统调用的开销很大，因此优化点落在了要求节点通信越过操作系统，直接在用户层进行实现通信。

因此，并行技术中我们需要不断探索低延迟高带宽的网络方案来提高效率。当今超级计算机在这方面是卓有成效的，如富岳的 Tofud 网络，美国 Frontier 的 Infinity Fabric 互联技术等等。

2.2.3 异构计算

富岳超算似乎没有使用 GPU，但纵览 TOP 500 榜单我们不难发现，美国 Frontier，summit 超级计算机均采用了 CPU+GPU 异构计算的模式 [4]，当 CPU 和 GPU 协同工作时，因为 CPU 包含几个专为串行处理而优化的核心，而 GPU 则由数以千计更小、更节能的核心组成，这些核心专为提供强劲的并行运算性能而设计，如图 2.4。程序的串行部分在 CPU 上运行，而并行部分则在 GPU 上运行，以此提高计算效率。这种计算模式主要应用在以下方面 [5]：

- 高性能计算领域，该架构可以增强算力和性能，目前受到广泛关注。

Figure 3 The GPU Devotes More Transistors to Data Processing

图 2.4: 异构模式

- 该架构适用于数据中心处理大量数据。
- 目前该架构的 CPU 和 GPU 内存池独立运算问题已经得到解决，可以共享内存空间，消除冗余内存，大大提高内存利用率，也因此越来越受到市场青睐。

3 中国的超算

3.1 中国超算的发展历史

一直以来，西方对中国采取科技封锁政策，这其中包括了超算，中国超算历史较于世界大国来说略显落后，中国国家层面的超级计算机自研事业始于 1978 年。那年 3 月在全国科学技术大会上，邓小平谈到“中国要搞四个现代化，不能没有巨型机”，开启了自研超级计算机的历史进程。但是中国凭借强大的科研实力，在超算方面发展迅速，跃升到国际先进水平国家中：

1. 1983 年中国成功研制出银河一号超级计算机，进而扩展银河系列超级计算机。
2. 1992 年中国成功研制出曙光一号超级计算机，此后开始进行向量型计算机到并行型计算机的发展转移。
3. 2002 年中国联想集团研制出深腾 1800 超级计算机，开始深腾系列超算的发展历程。
4. 2016 年中国发明出今天仍排在世界前列的神威 · 太湖之光超级计算机，其速度高达 9.3 亿亿次每秒，成为世界首台运行速度超十亿亿次的超级计算机。

3.2 神威 · 太湖之光中的并行架构

神威 · 太湖之光超级计算机由 40 个运算机柜和 8 个网络机柜组成。每个运算机柜中有 4 个由 32 块运算插件组成的超节点。每个插件由 4 个运算节点板组成，一个运算节点板又含 2 块“申威 26010”高性能处理器。一台机柜有 1024 块处理器，计算下来，整台“神威 · 太湖之光”共有 40960 块处理器，如图 3.5，可见多核集群在高性能计算方面的重要性 [6]。

申威 26010 处理器采用 64 位自主申威指令系统，峰值性能 3.168 万亿次每秒，核心工作频率 1.5GHz。后来又研发出 26010-pro 新处理器 [7]，SW26010-Pro 由 6 个核心组和 1 个协议处理单元构成，如图 3.6，每个核心组包含 64 个计算处理元素，总计 384 个内核，相比下 SW26010 只有 4 个核心组。它支持的内存控制器也从 DDR3 升级到 DDR4-3200；而且每个核心组都有自己的内存控制器，

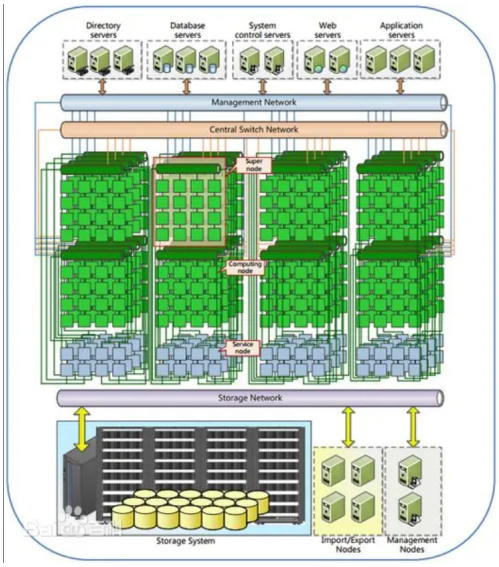

图 3.5: 神威 · 太湖之光架构图

配备 16 GB 内存，理论带宽达 307.2 GB/s。一个芯片能访问 96 GB 主内存，相比 SW26010 的 32 GB 实现了显著提升。可以说中国自主研发处理器的技术正在高质量发展。

在互联网络层面，神威采用硬软件组成的互联网络，其硬件主要由新一代神威路由器芯片和网络接口芯片构建而成。网络软件包括网络驱动、消息库、网络虚拟化、MPI、TCP/IP 和网络管理软件等。使得神威 太湖之光实现 1024 个处理器互连只使用了 4 层树网，提高了通信效率。

通过调研中外超级计算机中的架构体系，其实不难发现，核心仍然是针对处理单元集成的程度和互联通信效率的提升。

3.3 超算的对比

通过以上分析，神威 · 太湖之光和富岳超级计算机都具有较高的性能和并行计算能力。

神威 · 太湖之光的处理器核心数量更多，每个核心集成的计算能力更强，而富岳超级计算机的处理器则支持 SVE 指令集，能够进行高效的向量计算。

在通信网络方面，神威 · 太湖之光采用了三层自组织拓扑结构的软硬件综合网络，而富岳超级计算机采用了基于 6D Torus 结构的 TofuD 网络。这些网络都具有低延迟和高带宽的特点，能够满足大规模并行计算的通信需求。

总体而言，神威 · 太湖之光和富岳超级计算机在处理器和通信网络上都采用了先进的自主研发技术，为超级计算提供了强大的计算和通信能力。它们的差异主要体现在处理器核心数量和指令集支持上，以及通信网络的拓扑结构设计。

3.4 超算对于国家发展的意义

超级计算机属于战略高技术领域，是世界各国竞相角逐的科技制高点，也是一个国家科技实力的重要标志之一。在经历打压和封锁之后，中国超算经历了从无到有、从跟跑到局部领先、从关键核心技术引进到实现自主可控的艰难发展历程 [8]。研制超算可以让国家：

- 解除封锁，让高性能计算自主可控，不受制于人。

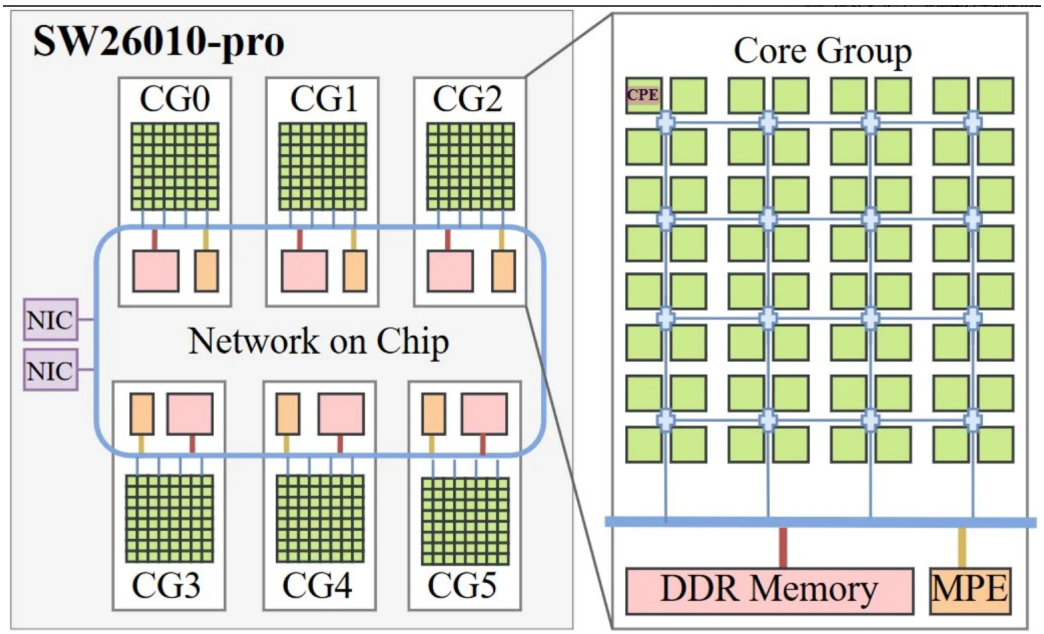

图 3.6: SW26010-Pro

- 加速应用，发挥“国之重器”的特殊作用，如天气气候、航空航天、船舶工程、海洋环境、石油勘探、生物信息、药物设计、电磁仿真、量子模拟、先进制造、新材料、新能源等多个应用领域。
- 继续推动科技创新、战略性新兴产业快速发展，推动大数据、人工智能、物联网快速发展。
- 在更宏大的层次上，为解决人类发展面临的重大挑战发挥重要作用。

4 并行体系结构对并行编程的意义

总的来说，并行体系结构的发展为编程提供了更高的计算能力，促进了新并行算法和数据结构的研究，也对程序员设计高效算法和可靠的代码提出了新要求，比如说：

- 多核 CPU 和异构 GPU 的迭代更新提供了多线程和计算过程的加速，如 CUDA 加速。
- MPI 是分布式内存系统的并行编程模型，可以很好的完成并行计算中的任务分配和消息传递任务。
- 随技术发展，编程方面也催生了多进程的并行快速排序（如图 4.7）和并行哈希表等利用并行体系结构的算法和数据结构。

5 调研总结

在本次并行调研中，我重点关注了超级计算机中的并行技术。并全面调研了以下几个方面的知识：

- 并行体系结构的概念。并行体系结构是指在计算机系统中具有并行性的体系结构，可以同时进行多个运算或操作。并行体系结构从时间和空间两个层次进行考虑，包括时间重叠和资源重复。

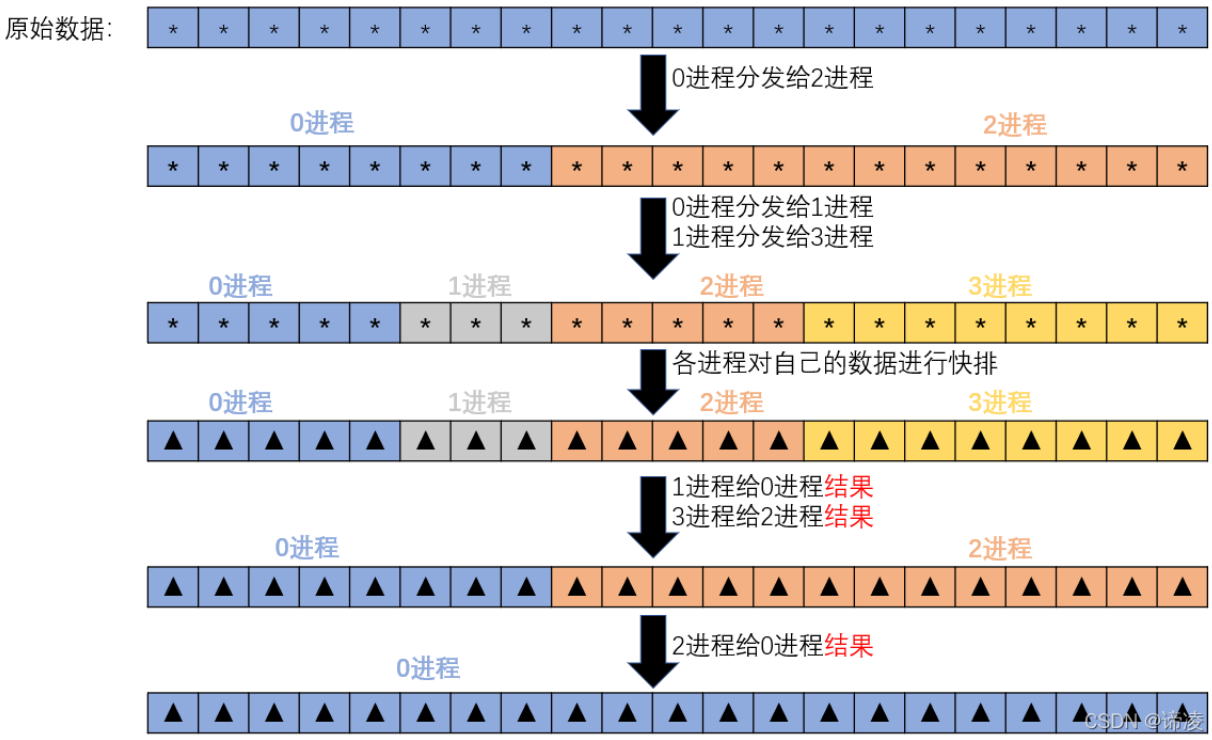

图 4.7: 并行快速排序

- 具体架构设计。并行体系结构的具体架构设计包括分布式系统、共享内存系统和异构系统(CPU+GPU)。关键技术点包括处理单元的组织、内存组织、通信方式等等。
- 超级计算机中的并行：我重点研究了富岳超级计算机系统和其基于 ARM 架构的超算集群。富岳超级计算机拥有大量的内核集群，其中使用 A64FX 核心处理器，具有出色的性能和支持 SVE 指令集的优点。该系统还采用了 TofuD 网络互连方案，提供高带宽内存直连和 6D Torus 网络接口。
- 中国的超算。我调研了中国超算的发展历史，并重点关注了神威 · 太湖之光超级计算机的并行架构。神威 · 太湖之光采用 SW26010 处理器，时至今日仍不断发展推出新一代高带宽高频率的 SW26010-Pro 处理器。而且中国的超算在国家发展中扮演着重要角色，对于科学研究和技术创新具有重要意义。
- 并行体系结构对编程的影响。并行编程需要显式地利用体系结构的并行性，并使用适当的并行编程模型、算法和数据结构来充分发挥并行体系结构的性能优势，两者相辅相成，相互赋能。

通过本次调研，我深入了解了并行体系结构和超级计算机中的并行技术。并行计算在解决复杂问题和提高计算效率方面具有重要作用，对于未来的计算机发展具有巨大潜力。

参考文献

[1] http://news.eeworld.com.cn/qrs/ic502000.html.

[2] https://www.cnblogs.com/kongchung/p/13184647.html.

[3] http://www.ict.ac.cn/liguojiewenxuan_162523/wzlj/lgjjs/201912/t20191227_5476630.html.

[4] https://www.chinastor.com/hpc-top500/0509406392019.html.

[5] https://zhuanlan.zhihu.com/p/615851976.

[6] https://baike.baidu.com/item/%E7%A5%9E%E5%A8%81%C2%B7%E5%A4%AA%E6%B9%96%E4%B9%8B%E5%85%89%E8%B6%85%E7%BA%A7%E8%AE%A1%E7%AE%97%E6%9C%BA/19755876.

[7] https://www.ithome.com/0/735/071.htm.

[8] https://news.cctv.com/2019/09/27/ARTIyzkjCpXLYOc3EXGhFQkE190927.shtml.