南開大學
Nankai University

计 算 机 学 院

并行程序设计实验报告

并行体系结构调研

张奥喆

年级：2023 级

专业：计算机科学与技术

指导教师：王刚

2025 年 3 月 15 日

摘要

进入人工智能时代后，传统串行计算难以支撑指数级增长的算力需求。本文梳理了中国超算的发展历史，并对各时期代表的我国并行体系结构进行分析，总结过去的发展。同时，本文通过对比美国 Frontier 超算，调研了目前我国并行架构相较于顶级超算的优势以及还需发展的方向。此外，本文还调查了未来并行体系可能的发展趋势，包括异构融合、智能调度、绿色计算等方向，阐明了自主架构的重要性。

关键字：并行体系结构；中国超算；未来趋势

目录

一、 我国超算的发展历史

(一) 起步阶段（1950 年代-1980 年代）：从零开始突破封锁

1. 早期计算机基础（1950-1970 年代）

1958 年，我国第一台通用数字电子计算机”103 机” 诞生，运算速度仅每秒 30 次，但标志着中国计算机事业的起点。1964 年，自主研发的”119 机” 实现每秒 5 万次浮点运算，首次用于东方红一号卫星轨道计算。

然而，20 世纪 70 年代，因国际技术封锁，我国被迫高价进口美国超算，并受限于” 玻璃房” 监控条款（美方控制机房和密码），如 CDC 公司设备需在透明监视下使用。

2. 自主研发巨型机的里程碑——银河系列

1978 年，邓小平提出” 中国要搞四个现代化，不能没有巨型机”，开启自主研制之路。1983 年，” 银河-I”（MIMD）亿次巨型机问世 [2]，使中国成为继美、日后第三个具备超算研制能力的国家。其完全国产化设计打破了封锁，为后续发展奠定基础。

(二) 快速发展阶段（1990 年代-2010 年代）：跻身世界前列

1. 曙光与天河系列的崛起

- 2001 年，” 曙光 3000” 突破每秒 4032 亿次运算，推动超算应用普及。
- **天河系列**：2009 年” 天河一号” 以 4700 万亿次/秒峰值速度首次登顶全球超算 TOP500 榜首（2010 年）[7]，并首创 CPU+GPU 异构架构，成为国际主流技术路线。这标志中国进入全球超算第一梯队。2013 年，” 天河二号” 实现六连冠，运算速度达 33.86 千万亿次/秒，

2. 全自主化的里程碑——神威系列

2016 年，” 神威·太湖之光” 以每秒 12.5 亿亿次运算登顶 TOP500，并**全部采用国产申威 26010 处理器**，首次实现硬件、软件完全自主化。其应用项目还斩获” 戈登贝尔奖”（高性能计算领域最高奖）[12]，彰显技术实力。

(三) 领跑阶段（2020 年代至今）：迈向百亿亿次与智能化

E 级超算的突破：2021 年，” 天河” 新一代百亿亿次系统（E 级）完成部署，算力达每秒超百亿亿次，支撑人工智能、量子计算等前沿领域。

二、 我国超算的并行体系的发展和分析

下面将分析三个我国的代表超算，他们可以代表各自时期中国并行体系结构，通过对于这些超算的分析，可以调研出中国并行体系结构的发展。

(一) 银河-I（MIMD）[10]

1. 性能

每秒 1 亿次浮点运算（峰值）

2. 硬件架构：向量处理与模块化设计

1. **向量处理单元（VPU）**
   银河-I 采用**向量运算为主**的架构，字长 64 位，支持大规模科学计算中的向量化操作。其核心创新在于**双向量阵列结构**：

   - **双阵列并行**：设置两组独立的向量寄存器组，允许同时处理多个向量数据流，通过交叉存取减少数据冲突，提升吞吐量。
   - **流水线技术**：借鉴 Cray-1 的流水线设计，将浮点运算分解为多级流水，实现指令级并行（ILP），单个运算单元可连续处理多个向量元素。

2. **分布式处理与模块化布局**

   - 主机系统由**并行处理器**（时钟频率 20MHz）和**浮点运算器**（主频 40MHz）构成，两者协同工作，形成任务级并行（TLP）能力。
   - 硬件采用**模块化设计**，包含多层印刷电路板和高密度组装技术，便于维护和扩展。例如，存储系统分为磁芯主存（16MB）和磁盘外存（256MB），支持层次化数据访问。

3. **并行策略：层次化并行与资源调度**

   - **指令级并行（ILP）**
     - 通过流水线技术和向量寄存器重用，单条指令可并行处理多个数据元素。例如，向量加法指令可一次性完成多个浮点数的叠加。
   - **数据级并行（DLP）**
     - **双向量阵列**允许同时处理两组向量数据，例如在气象模拟中，可并行计算不同区域的温度场和压力场。
     - 存储系统通过**交叉编址技术**实现多体并行访问，缓解内存带宽瓶颈。
   - **任务级并行（TLP）**
     - 操作系统支持多任务调度，可将大型问题分解为子任务分配给不同处理单元。例如，在核武器模拟中，物理方程的求解与边界条件处理可并行执行。

3. 软件生态适配

- 开发了兼容国际标准的编译系统，支持**向量化编程**和并行算法库，用户可通过高级语言调用并行计算资源。
- 系统软件实现了任务调度与资源分配的动态优化，减少并行任务间的资源竞争。

4. 局限

- **并行粒度受限**：受制于 1980 年代工艺水平，处理器数量和互联带宽有限，难以实现更高维度的并行扩展。
- **编程复杂度高**：用户需手动优化向量化代码，与当代自动并行化工具相比门槛较高。

(二) 天河一号 [9]

1. 性能

每秒 4700 万亿次（4700TFlop/s）运算

2. 硬件架构：异构计算（首创）

1. 异构计算单元配置

- **CPU-GPU 协同架构**：系统采用 Intel Xeon CPU 与 NVIDIA Tesla GPU 的混合架构，其中 CPU 负责逻辑控制与通用计算，GPU 专注于高并行浮点运算。例如，TH-1A 版本包含 14336 个 Intel Xeon X5670 六核 CPU 和 7168 个 NVIDIA Tesla M2050 GPU，总计算核心超过 20 万。
- **自主处理器集成**：二期系统（TH-1A）引入国产飞腾 FT-1000 八核处理器（主频 1GHz），用于服务节点的管理任务，标志着国产芯片在高性能计算中的突破。

异构计算：目前主流的并行结构

异构计算（Heterogeneous Computing）是一种利用多种不同类型的处理器或计算单元协同工作的计算模式，旨在通过不同架构的互补优势，提升系统的整体性能、能效或灵活性。传统计算通常依赖单一类型的处理器（如 CPU），而异构计算通过整合 CPU、GPU、FPGA、ASIC 等不同架构的硬件，针对特定任务选择最合适的计算单元，从而实现更高效的计算。异构计算已成为当前并行架构发展的核心趋势。

1. 硬件

异构系统包含多种处理器类型，例如：

- **CPU（中央处理器）**：擅长复杂逻辑控制和串行任务。
- **GPU（图形处理器）**：擅长高并行计算（如图形渲染、深度学习）。
- **FPGA（现场可编程门阵列）**：可灵活配置硬件逻辑，适合定制化加速。
- **ASIC（专用集成电路）**：针对特定任务优化（如 TPU 用于 AI 推理）。
- **AI 加速器（如 NPU）**：专为神经网络设计的计算单元。

2. 软件（编程模型）

- **OpenCL (Open Computing Language)**
- **SYCL (C++ Single-source Heterogeneous Programming)**
- **CUDA** NVIDIA GPU 专用编程框架。优势在于性能优化极致，生态成熟（如 cuDNN、cuBLAS 等加速库）。处于行业垄断地位。

2. 高速互连通信系统

- **定制化网络芯片**：采用自主研发的 NRC 路由芯片和 NIC 接口芯片，构建光电混合的胖树拓扑网络，链路双向带宽达 160Gbps（较初期 40Gbps 提升 4 倍），单背板交换密度 61.44Tbps，显著优于同期商用互连技术。
- **低延迟优化**：通信延迟低至 1.2 微秒（初期版本为 1.57 微秒），满足数万节点间的高效数据交换需求。

3. 层次化存储与 I/O 设计

- **大容量内存与存储**：一期系统配备 98TB 内存和 1PB 共享磁盘；升级后的 TH-1A 扩展至 262TB 内存和 2PB 存储，支持 Lustre 并行文件系统，实现 PB 级数据的全局共享访问。

- **分布式存储结构**：通过 128 个对象存储节点和 6 个元数据节点优化 I/O 性能，提升大规模数据读写效率。

4. 高密度组装与能效控制

- **双面对插机柜布局**：通过高密度集成降低物理空间占用，全系统 140 个机柜占地仅 700 平方米。
- **动态功耗管理**：结合处理器调频调压、自适应节点能耗状态转换等技术，系统总功耗控制在 4.04MW，能效比达 431.7MFlops/W。

3. 软件系统

天河一号的软件体系围绕**异构协同编程**和**高效资源管理**展开，核心技术包括：

1. 定制化操作系统

- **麒麟 Linux 优化版**：基于 64 位 Linux 内核的银河麒麟系统，针对高性能计算需求强化了能耗管理、虚拟化安全域（支持用户环境隔离）及并行任务调度功能，符合 B2 级安全标准。

2. 异构编程与编译支持

- **多语言与并行框架**：支持 C/C++、Fortran、Java 等语言，集成 OpenMP、MPI 并行编程模型，并提供 **TH-HPI 异构编程接口**，简化 CPU 与 GPU 的协同开发流程。
- **编译优化技术**：采用多级并行动态负载均衡算法和全程序过程间分析，提升异构资源利用率，使 GPU 计算效率从理论值的 20% 提升至 70%。

3. 资源管理与容错机制

- **虚拟计算域技术**：通过容器化实现用户环境的安全隔离与定制，支持多任务并行运行（如 TH-1A 每日处理超 1400 个作业）。
- **多层次容错设计**：硬件级监控诊断与软件级容错结合，保障数万节点长时间稳定运行，故障恢复时间缩短至分钟级。

4. 应用支撑与工具生态

- **并行开发环境（FSE）**：提供集成调试工具、任务调度器及可视化平台，覆盖从代码开发到结果分析的全流程。
- **领域专用优化**：针对气象预报、基因测序等应用优化算法，例如通过基因数据分析实现无创胎儿健康筛查，日均处理数据量达 PB 级。

（三） 神威·太湖之光 [11]

1. 性能

每秒 12.5 亿亿次（125PFlop/s）运算

2. 硬件并行架构：异构众核与高密度扩展

1. 申威 26010 众核处理器（自主研发）

   - **芯片级并行**：每个处理器采用**异构众核架构**，包含 4 个运算控制核心（MPE）和 64 个运算核心（CPE），总计 260 个核心。CPE 阵列以” 单指令快多数据流”（SBMD）模型运行，相比传统 SIMD 减少内存访问延迟，比 MIMD 提升指令执行效率。
   - **片上存储优化**：集成 32KB 指令缓存和 64KB 用户可控本地存储（LDM），通过数据预取和寄存器通信降低核间通信开销。

2. 系统级扩展性

   - **超节点模块化设计**：全系统由 40 个运算机柜构成，每个机柜含 1024 个处理器，总计 40960 个处理器。通过高密度弹性超节点布局，支持从单节点到千万核规模的灵活扩展。
   - **复合网络互连**：采用高流量定制网络架构，对分带宽达 70TB/s，确保超大规模节点间低延迟、高吞吐通信。

3. 存储与通信并行策略

   - **在线存储系统**：288 台 SSD 存储节点提供 341GB/s 聚合带宽，支持高并发 I/O 访问；近线存储系统通过 SAN 实现大容量冷数据存储。
   - **多级缓存机制**：结合 SSD 和内存缓存层，采用动态数据预取与分布式缓存一致性协议，提升数据局部性。
   - **自适应路由算法**：根据网络负载动态调整数据传输路径，避免热点阻塞。
   - **主从核协同通信**：MPE 负责全局任务调度，CPE 通过片上网络直接交换数据，减少主核干预 [1]。

3. 软件生态与并行编程模型

1. 并行编程框架

   - **两级并行化策略**：任务级（MPI 进程间通信）与数据级（CPE 线程并行）结合，支持千万核规模任务分解。
   - **自主编译工具链**：支持 OpenACC、OpenMP 等并行标准，并提供神威定制指令集优化库（如 SWMath），实现算法与硬件的深度适配。
   - 全系统采用自主指令集，摆脱对 X86/GPU 架构依赖，实现从芯片到系统的全栈可控。

2. 容错与调度机制

   - **主动容错体系**：通过硬件冗余与软件检查点技术，实现错误检测与自动恢复，单晚可完成三次全系统 Linpack 测试。
   - **动态负载均衡**：基于任务类型与资源状态的自适应调度算法，优化计算密集型与数据密集型任务的资源分配。

三、 进阶要求

(一) Frontier（美国）[5]

1. 性能

2023 年 TOP500 榜首（1.194 EFLOP/S）

2. 硬件架构

- **异构计算架构：**每个节点带有 1 个高性能和 AI 计算负载优化的 AMD EPYC 7453s 64 核 CPU 和 4 个 AMD Radeon Instinct MI250X GPU 整套系统包括 9,472 个 CPU 和 37,888 个 GPU，总计 CPU 内核数达 606,208 个，GPU 内核数达 8,335,360 个 [5]
- **互连网络**：Cray Slingshot-11 网络拓扑，200 Gb/s 带宽，采用自适应路由和拥塞控制技术。
- **节点内并行**：CPU 与 GPU 共享内存空间，支持统一寻址（HSA 架构）。
- **内存与存储**：每个计算节点配备高速内存，以满足大规模数据集的快速访问需求。此外，Frontier 采用并行文件系统（如 Lustre 或 GPFS）和大规模存储阵列，提供 PB 级甚至 EB 级的海量存储容量，支持高吞吐量、低延迟的 I/O 需求。
- 能效比达 52.23 GFLOPS/W（Green500 排名前列）。全系统采用慧与 Cray EX 液冷机柜，结合室温水冷却方案，冷却能耗仅占总功耗的 3%-4%（相比前代 Summit 的 10% 大幅优化）[6]。

3. 软件与应用生态

- 操作系统：Frontier 采用 HPE 子公司克雷公司开发的 Cray OS 作为操作系统，确保系统的稳定性和高效性。
- 编程模型与工具：Frontier 支持多种并行编程模型，如 MPI 和 OpenMP，研究人员将 Megatron-DeepSpeed 分布式训练框架移植到 Frontier 上，以支持在 AMD 硬件和 ROCm 软件平台上进行高效的分布式训练。

(二) Frontier 与我国超算对比分析

1. 细节对比

由于我国目前更新推出来的超算没有参与 Top500 排名，技术细节也没有公开（战略保密），因此这里拿神威·太湖之光来进行比较。

维度	Frontier (美国)	神威太湖之光 (中国)
处理器架构	AMD EPYC + Instinct GPU	申威 26010 众核处理器（260 核/片）
并行互连	Slingshot-11 (商用)	自主 SW 互连网络（定制协议）
峰值算力	1.69 EFLOPS	125 PFLOPS (待续)
能效比	51.5 GFlops/Watt	6.05 GFlops/Watt
软件生态	ROCm+OpenMP/OpenACC	自主研发编译器 + 神威框架

表 1: 超级计算机比较

2. 我国超算的劣势与发展方向

1. 软件生态薄弱 [3]

   - 应用适配困难：国产超算（如神威、天河）采用自主指令集，需重构 90% 以上传统软件代码。Frontier 支持 HIP/ROCm 生态系统 [5]，而国产框架开发者适配率不足 30%。
   - 商业软件依赖：制造业、气象领域仍依赖 ANSYS、WRF 等国外软件，年采购费用超 20 亿元，且升级受制于外方协议条款。

2. 硬件制程差距

   - 芯片代际鸿沟：申威处理器制程仍停留在 14nm（神威·太湖之光升级版），而美国 Frontier 超算已采用 6nm 工艺，单核性能差距达 3 倍 [4]。
   - 电力成本高：单台 E 级超算年耗电超 2 亿度（如天河三号），冷却系统占运维成本 40%。对比美国 Frontier 液冷技术，国产超算 PUE 值普遍高于 1.2 [8]，能效优化空间显著。

（三） 并行体系结构的未来特征

未来并行架构将呈现异构化、智能化、绿色化、安全化的特征，深度融合硬件创新与软件生态，服务于 AI、边缘计算、量子模拟等新兴领域。

趋势方向	技术方向	具体趋势	示例/技术	应用场景
标准化与开放生态	统一接口与路径、平台兼容	推动 OpenMP、MPI 等标准化接口,解决异构硬件互操作性	RISC-V 指令集、向量扩展	异构计算平台开发（CPU/GPU/FPGA 协同）
	开源框架与容器化	优化分布式框架资源调度,结合 Kubernetes 实现云边端协同	Spark、Hadoop、Kubernetes	云计算、边缘计算资源动态分配
异构计算与能效优化	混合架构设计	CPU+GPU+FPGA/TPU 组合，兼顾高性能与低延迟	NVIDIA Grace Hopper 超算架构	科学计算、AI 推理与训练
	绿色计算技术	动态电压频率调整（DVFS）、能耗预测算法、碳基芯片材料	碳基芯片、液冷散热系统	数据中心、超算中心能效优化
智能化与 AI 驱动	AI 原生并行设计	自动任务划分、动态负载均衡，减少人工调参	TensorFlow/PyTorch 分布式编译器	大规模深度学习模型训练
	自适应资源管理	基于强化学习的动态资源调度系统	Google Borg、Kubernetes Autoscaler	云计算、实时数据处理任务调度
安全与隐私	内生安全机制	内置差分隐私、同态加密模块，保障数据安全	联邦学习框架（如 FATE）	医疗数据共享、跨企业协作
	容错能力增强	冗余节点部署、实时错误检测（TMR 三模冗余）	航天仿真系统、核反应堆模拟	高可靠性要求的科学工程领域

表 2: 计算技术趋势与应用

四、 总结

通过本次调研可以看出来，我国超算经历了从无到有，从落后到顶尖的过程。纵观发展历史，我国超算在银河、天河、神威等机型中逐渐构建了具有自主架构的体系结构。当前我国虽然已经实现了 E 级的算力突破，但是在软件、芯片制程等方面还是存在明显的短板。这启示我们，未来并行体系需要硬件创新也需要软件生态构建。特别是在 AI 大模型时代, 突破 CUDA 等垄断壁垒，探索智能资源调度等前沿方向，推动并行计算向绿色化，是未来我国未来并行架构需要重点研究的方向。

参考文献

[1] H. Fu et al. Sw26010 众核处理器访存优化. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC'16)*, pages 26:1–26:12. IEEE Press, 2016.

[2] J. L. Hennessy and D. A. Patterson. **计算机体系结构：量化研究方法**. 机械工业出版社, 7 edition, 2017.

[3] Hyperion Research. Hpc market trends report. Technical report, Hyperion Research, 2022.

[4] IEEE Spectrum. 先进制程技术路线图. *IEEE Spectrum*, 2022.

[5] Oak Ridge National Laboratory. Frontier 应用生态报告. Technical report, Oak Ridge National Laboratory, 2022.

[6] Oak Ridge National Laboratory. Frontier 超算液冷系统设计. In *Proceedings of the IEEE International Symposium on High-Performance Computer Architecture (HPCA'23)*, pages 123–134. IEEE, 2023.

[7] TOP500. Top500 official list. https://www.top500.org, 2023.

[8] 中国电子技术标准化院. 超算能效测试规范. Technical report, 中国电子技术标准化院, 2021.

[9] 国家超算天津中心. 天河系统架构白皮书. Technical report, 国家超算天津中心, 2020.

[10] 国防科技大学. 银河系列超算技术报告. Technical report, 国防科技大学, 2020.

[11] 江南计算所. **申威** *26010* **处理器技术手册**. 江南计算所, 2016.

[12] 清华大学. 戈登贝尔奖获奖项目技术说明. Technical report, 清华大学, 2021.