并行计算与大数据处理技术简介

为什么要并行计算

提高计算机性能

单核性能逐渐达到极限（ILP wall，memory wall，etc.），所以需要并行计算

应用领域计算规模和复杂度上升

并行计算技术分类

按数据和指令处理结构：Flynn 分类

	single instruction	**multiple instruction**
single data	SISD	MISD
multiple data	SIMD	MIMD

按并行类型分类：位级，指令级，线程级（数据级并行&任务级并行）

按访存结构分类：共享内存，分布式共享存储，分布式内存

按系统类型分类：多核并行计算（MC），对称多处理系统（SMP），大规模并行处理（MPP），集群，网格

按计算特征分类：数据密集型，计算密集型

按程序设计模型分类：共享内存变量，消息传递（MP），**MapReduce**方式

主要技术问题

多核/多处理器的互联

访存体系结构

分布式数据与文件管理

并行计算算法设计

并行计算程序模型

数据的同步和通信控制

可靠性设计

并行计算的软件框架平台

系统性能评估和并行度评估

> Amdahl 定律
>
> $$S = \frac{1}{(1-P) + \frac{P}{N}}$$
>
> S 为加速比，P 为程序中可并行的比例，N 为处理器个数

MPI 并行程序设计

MPI (Message Passing Interface)，基于消息传递的高性能并行计算编程接口

为处理器间提供可靠的、面向消息的通信

主要功能

- 点对点通信：同步/异步
- 节点集合通信：广播，同步控制，对结果的归约
- 支持对自定义复合数据类型传输