MPP 大规模并行处理机 和 Cluster

MPP概述

MPP（Massively Parallel Processing） 是一种基于**分布式计算**的体系结构，旨在通过大量独立的计算节点并行执行计算任务。其核心思想是将一个大的数据处理任务分解为多个小的子任务，并在多个处理器上同时执行这些子任务，以实现高效的并行处理。每个节点都有独立的磁盘存储系统和内存系统，业务数据根据数据库模型和应用特点划分到各个节点上，每台数据节点通过专用网络或者商业通用网络互相连接，彼此协同计算。MPP 具有完全的可伸缩性、高可用、高性能、优秀的性价比、资源共享等优势。

节点互联机制 不同于NUMA 是在单台物理服务器内部，通过高速互联实现处理器与本地内存的非一致性访问，MPP 是通过网络互联实现多个节点的分布式计算，每个节点通常是一个独立的 SMP 系统。节点互联通常由 **专用硬件** 和 **高速通信网络** 组成。这些计算节点通常拥有独立的内存，但节点间的通信延迟较低，通常采用 **专用的高速网络**（如 InfiniBand）来减少通信开销。

- 特点：
  - **任务并行执行：**MPP架构能够将一个大的数据处理任务分解为多个小的子任务，并在多个节点上并行执行这些子任务。这种并行处理方式能够显著提高数据处理速度。
  - **数据分布式存储：**在MPP架构中，数据被分散存储在各个节点上。这种数据分布式存储方式使得每个节点都能独立处理数据，并与其他节点协同工作。
  - **分布式计算：**MPP架构采用分布式计算方式，每个节点独立计算一部分数据，并将结果汇总得到最终结果。这种方式能够提高计算效率，并减少单点故障的风险。
  - **高并发：**MPP架构具有高并发性，单个节点能够支持大于300用户的并发访问。这种高并发性使得MPP架构能够高效地处理大规模数据集。
  - **横向扩展：**MPP架构支持集群节点的扩容，随着数据规模的增加，可以通过增加节点数量来扩展处理能力。这种横向扩展方式使得MPP架构具有良好的可扩展性。
  - **Shared Nothing（完全无共享）架构：**在MPP架构中，每个节点拥有自己的CPU、内存和其他硬件资源，互不共享。这种Shared Nothing架构使得MPP架构具有很好的稳定性和可靠性。

提出时间与代表性论文

- **提出时间**：1970s-1980s
- 代表性论文：
  - **Hennesy, J. L., & Patterson, D. A. (1990). *Computer Architecture: A Quantitative Approach***：尽管这本书没有单独介绍 MPP，但它提供了现代多处理器架构的基础概念，帮助理解多处理器并行计算模型。
  - **Kirk, D. (1983). "The Design of a Highly Parallel Processor"**：这篇文章描述了早期的 MPP 设计理念。

现实世界中的机器例子

MPP 系统通常使用高性能的计算机集群，具有专用硬件和优化的通信系统。以下是一些典型的 MPP 系统：

1. **Cray T3E**（1990s，Top500 超级计算机榜单上出现）

    - **Cray T3E** 是 Cray 公司推出的一款典型的 MPP 超级计算机，采用了大量的并行处理单元，并支持大规模的并行计算。
    - **性能**：Cray T3E 的峰值性能为 **1.5 TFLOPS**。
2. **IBM Blue Gene/P**（2000s，曾位居 Top500）
    - **IBM Blue Gene/P** 是一款典型的 MPP 超级计算机，拥有大量并行处理单元，设计用于超大规模并行计算。
    - **性能**：其峰值性能达到 **1.02 PFLOPS**，在 2008 年的 Top500 排名中名列第一。
3. **Fujitsu K Computer**（2011年，Top500 排名）
    - **Fujitsu K** 是另一款代表性的 MPP 超级计算机，用于各种科学研究。
    - **性能**：其峰值性能达到 **10.5 PFLOPS**，曾在 Top500 榜单上位居第一。
4. **NVIDIA DGX SuperPOD**
    - 基于 **NVIDIA Tesla GPU** 的 MPP 设计，尤其适用于深度学习和人工智能训练。
    - **性能**：每个 **DGX SuperPOD** 集群的峰值性能可达到 **500 PFLOPS** 以上。

机群Cluster

计算集群通常指的是由多台独立的计算机（节点）组成的一个集合。这些计算机通过网络互联，共享任务来完成并行计算。计算集群的计算资源可以是同质的（所有节点具有相同硬件）或异质的（不同节点可能具有不同硬件配置）。集群系统通常使用 **分布式内存** 或 **共享内存**（取决于集群的实现）。集群中的计算节点可以是低成本的标准服务器，通常配置为运行 Linux、Windows 或其他操作系统。

节点互联集群系统通常依赖于**标准硬件**，而不是专用硬件。资源可能分散在多个物理机上，这些物理机通过标准网络连接（如千兆以太网、万兆以太网或InfiniBand）进行通信。集群的资源管理通常依赖于**操作系统**和**调度软件**（如 Slurm、PBS、Torque、Hadoop 等）。计算集群的通信可能不像 MPP 那样具备专门的硬件支持，通信延迟相对较高，特别是在大规模集群中。

- **优点**
    - 成本低：基于普通硬件，性价比高。
    - 灵活性高：可轻松扩展或替换节点。
    - 容错性：单个节点失效时不会导致整个系统瘫痪。
- **缺点**：
    - **管理复杂性**：集群的管理和维护相对复杂，尤其是在大规模集群中。需要处理节点的故障、负载均衡、资源调度等任务。
    - **通信开销**：集群中节点之间的通信依赖于网络带宽和延迟，在大规模集群中，通信开销可能会成为瓶颈。
    - **一致性问题**：在分布式存储系统中，如何保证数据的一致性和完整性是一个重要的问题，尤其是在节点之间频繁进行数据交换时

提出时间与代表性论文

- **提出时间**：1990s 及之后
- 代表性论文：

- **Buyya, R., & Murshed, M. (2002). "GridSim: A Toolkit for the Modeling and Simulation of Grid Computing Environments"**：该论文介绍了 **GridSim** 工具包，这是为计算集群和网格计算环境建模与模拟而设计的工具。
- **Culler, D. E., & Singh, J. P. (1999). "Parallel Computer Architecture: A Hardware/Software Approach"**：本书描述了并行计算的基本架构和集群系统的原理。

现实世界中的机器例子

计算集群的广泛应用使得其成为现代数据中心和超级计算机的基础。以下是一些典型的集群实例：

1. **Top500 中的许多系统**：
   - Top500排行榜上很多位于前列的计算机实际上都是

     集群系统，例如：
     - Sunway TaihuLight（中国，2016 年）
     - **性能**：93 PFLOPS，虽然它的架构采用了 **定制的处理器**，但其工作原理是基于大量的处理单元和高效的通信网络，这使其属于 **集群类型**。
2. **Alibaba Cloud** 和 **Amazon Web Services (AWS)** 的大规模计算集群：
   - 这些云计算平台为客户提供的计算资源实际上是基于大量标准服务器的计算集群，用于处理数据分析、AI、深度学习等任务。
3. **Google** 的 **TensorFlow Processing Units (TPU)** 集群：
   - Google Cloud 上的 TPU 集群采用专为机器学习优化的处理单元，分布式计算资源组成的集群支持深度学习等应用。
4. **Microsoft Azure** 和 **IBM Cloud**：这些云平台通过 **计算集群** 提供弹性计算资源，满足大规模数据处理和高性能计算需求。

两者对比

特性	MPP	计算集群
架构	大规模并行计算专用硬件，紧密集成的节点，优化通信	由多台独立计算机（节点）组成，资源可以共享或独立
资源管理	专用硬件，优化的资源调度和任务调度	标准硬件，灵活的资源调度和管理工具（如Slurm、PBS等）
通信方式	高效的高速网络（如InfiniBand）连接，低延迟通信	通过标准网络连接，通信开销可能更大，依赖于调度软件
扩展性	有限，增加节点时可能需要专门的硬件资源支持	良好，扩展非常灵活，增加节点较为容易
应用场景	高性能计算、科学模拟、工程计算等	大数据处理、Web服务、虚拟化、大规模并行任务等

两者的区别在于，**MPP** 系统往往设计上更加集成、专用，追求极致的并行计算性能，而 **集群系统** 则更多依赖通用计算节点，通过灵活的资源管理实现分布式计算，适用范围更广。

参考文献

1. Kirk, D. (1983). *The Design of a Highly Parallel Processor*. In *Proceedings of the 16th Annual International Symposium on Computer Architecture* (pp. 227-235). IEEE.
2. Hennessy, J. L., & Patterson, D. A. (1990). *Computer Architecture: A Quantitative Approach* (2nd ed.). Morgan Kaufmann Publishers.
3. Buyya, R., & Murshed, M. (2002). *GridSim: A Toolkit for the Modeling and Simulation of Grid Computing Environments*. *Software: Practice and Experience*, 32(2), 1-22.
4. Culler, D. E., & Singh, J. P. (1999). *Parallel Computer Architecture: A Hardware/Software Approach*. Morgan Kaufmann Publishers.
5. Dongarra, J. J., & Kogge, P. M. (2001). *The Top500 Supercomputing Sites*. *Journal of Supercomputing*, 13(1), 19-26.
6. **Sunway TaihuLight (2016).** *China's Sunway TaihuLight: The World's Fastest Supercomputer*. *Top500 Supercomputing Sites*. Retrieved from https://www.top500.org
7. **Fujitsu K Computer (2011).** *Fujitsu K Computer: Japan's Most Powerful Supercomputer*. *Top500 Supercomputing Sites*. Retrieved from https://www.top500.org