

КАЗАХСКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ ИМЕНИ АЛЬ-ФАРАБИ

А.Т. Агишев

ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ ОБРАБОТКИ ДАННЫХ

**Сборник лекций для студентов магистратуры,
обучающихся по образовательной программе
«7М07125 - Электроника и системы управления»**

Алматы, 2025

Лекция 8. Внутреннее устройство серверных нод и принципы их масштабирования

Цель лекции

Рассмотреть структуру и принципы функционирования серверных нод в вычислительных системах. Изучить организацию аппаратных и программных компонентов, способы увеличения вычислительной мощности и подходы к масштабированию серверных узлов в составе кластеров.

Основные вопросы:

1. Понятие и назначение серверной ноды.
2. Архитектура и основные компоненты узла.
3. Многоядерные и многопроцессорные конфигурации.
4. Организация памяти и взаимодействие с подсистемой ввода-вывода.
5. Взаимодействие процессоров и шинные интерфейсы.
6. Принципы масштабирования: вертикальное и горизонтальное.
7. Проблемы масштабируемости и методы их решения

Краткие тезисы:

1. Понятие серверной ноды. Серверная нода — это базовый элемент вычислительного кластера, включающий процессоры, оперативную память, систему хранения и сетевые интерфейсы. Каждая нода выполняет часть общей задачи, взаимодействуя с другими через высокоскоростную сеть. Эффективность всей системы зависит от баланса производительности вычислений, пропускной способности памяти и скорости межузлового обмена.

2. Архитектура и компоненты узла. Современные серверные ноды включают:

- **Центральные процессоры (CPU):** многоядерные чипы с поддержкой многопоточности и SIMD-инструкций;
- **Оперативную память (RAM):** модульная система DDR4/DDR5 с высокой пропускной способностью;
- **Систему хранения данных:** NVMe SSD, RAID-массивы или подключение к общему хранилищу;
- **Сетевые адаптеры:** Ethernet, InfiniBand или специализированные интерфейсы;
- **Систему охлаждения и питания:** обеспечивают стабильность работы при нагрузке;
- **Контроллер управления (BMC/IPMI):** для удалённого мониторинга и администрирования.

Конфигурация узла определяется его ролью — вычислительной, графической или хранилищной.

3. Многоядерные и многопроцессорные системы. Современные серверные узлы используют процессоры с десятками или сотнями ядер (например, AMD EPYC, Intel Xeon).

- **Многоядерность** позволяет параллельно выполнять множество потоков внутри одного процессора.
- **Многопроцессорные конфигурации (SMP, NUMA)** объединяют несколько процессоров, разделяющих память и ресурсы.

При этом особое внимание уделяется организации памяти, чтобы минимизировать задержки при доступе к удалённым блокам (NUMA-оптимизация).

4. Подсистема памяти и ввод-вывод. Производительность узла во многом определяется скоростью обмена данными между CPU, памятью и устройствами ввода-вывода. Основные интерфейсы:

- **PCI Express (PCIe):** соединяет процессоры с ускорителями и накопителями;
- **DDR4/DDR5:** обеспечивают высокую скорость передачи данных (до сотен ГБ/с);
- **NVLink, Infinity Fabric:** шины для взаимодействия между процессорами и GPU.

Кэширование и буферизация данных позволяют снижать задержки, а многоканальные контроллеры памяти обеспечивают равномерную загрузку ресурсов.

5. Взаимодействие процессоров и шинные интерфейсы. Обмен между процессорами осуществляется по скоростным межсоединениям: QPI, UPI, Infinity Fabric, NVSwitch. При увеличении количества процессоров важно сохранять согласованность кэша (cache coherence) и минимизировать задержки доступа к данным. Для этого применяются протоколы когерентности и иерархические контроллеры памяти.

6. Принципы масштабирования. Масштабирование серверных систем делится на два типа:

- **Вертикальное масштабирование (scale-up):** увеличение ресурсов одного узла — добавление процессоров, памяти, ускорителей. Преимущества — простота управления, недостатки — ограниченная масштабируемость и высокая стоимость.
- **Горизонтальное масштабирование (scale-out):** добавление новых узлов в кластер. Обеспечивает гибкость и отказоустойчивость, но требует развитых сетевых технологий и систем управления задачами.

Большинство современных НРС и облачных систем используют именно горизонтальное масштабирование.

7. Проблемы масштабируемости и пути их решения. Основные ограничения при увеличении числа узлов:

- рост сетевых задержек;
- увеличение накладных расходов на синхронизацию;
- ограниченная пропускная способность шины и памяти;
- энергетические и тепловые ограничения.

Для повышения эффективности применяются:

- балансировка нагрузки по узлам;
- оптимизация алгоритмов под параллельное исполнение;
- использование GPU и специализированных ускорителей;
- внедрение энергоэффективных решений и интеллектуальных систем мониторинга.

Вопросы для контроля, изучаемого материал:

- 1) Что представляет собой серверная нода и какие компоненты она включает?
- 2) В чём различие между многоядерной и многопроцессорной архитектурой?
- 3) Как организовано взаимодействие процессоров и подсистемы памяти?
- 4) В чём различие между вертикальным и горизонтальным масштабированием?
- 5) Какие факторы ограничивают масштабируемость серверных систем?
- 6) Как современные технологии позволяют повышать эффективность масштабируемых вычислений?

Рекомендуемый список литературных источников:

1. Hopgood A. A. Intelligent Systems for Engineers and Scientists. — 3rd ed. — Boca Raton : CRC Press / Taylor & Francis, 2012. — 682 p.
2. Sterling T., Anderson M., Brodowicz M. High Performance Computing: Modern Systems and Practices. — Amsterdam : Elsevier / Morgan Kaufmann, 2017. — 728 p.
3. Russell, S., Norvig, P. Artificial Intelligence: A Modern Approach. 4th Edition. Pearson, 2021.