

Лекция 1. Организация функционирования распределённых вычислительных систем

Кулагин Иван Иванович

ст. преп. Кафедры вычислительных систем
Сибирский государственный университет
телекоммуникаций и информатики

Created by:

Пазников Алексей Александрович
к.т.н. доцент Кафедры вычислительных систем



Объект курса?

Распределённые вычислительные системы

Предмет курса?

*Модели и алгоритмы организации
функционирования*



Что хочет пользователь?

- Сокращение времени решения задач.
- Решение задач, требующих огромных объёмов памяти.

В то же время

совершенствование средств ВТ на основе модели вычислителя *не даст кардинального улучшения технических характеристик.*



Распределённая ВС — мультипроцессорные ВС с MIMD-архитектурой, в которых нет единого ресурса.

Представляется множеством взаимодействующих элементарных машин, оснащенных средствами коммуникаций и внешними устройствами.

Архитектурные особенности

- Иерархическая структура
- Мультиархитектурная организация
- Разнородность состава



Список TOP500 (июнь 2013)

	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	National University of Defense Technology China	Tianhe-2 (MilkyWay-2) - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
2	DOE/SC/Oak Ridge National Laboratory United States	Titan - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209
3	DOE/NNSA/LLNL United States	Sequoia - BlueGene/Q , Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890
4	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer , SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705,024	10,510.0	11,280.4	12,660
5	DOE/SC/Argonne National Laboratory United States	Mira - BlueGene/Q , Power BQC 16C 1.60GHz, Custom IBM	786,432	8,586.6	10,066.3	3,945



Tianhe-2 (MilkyWay-2) - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P

Расположение:	Национальный университет оборонных технологий (Китай)
Производитель:	NUDT
Количество ядер:	3,120,000
Производительность Linpack (Rmax)	33,862.7 teraFLOPS
Пиковая производительность (Rpeak)	54,902.4 teraFLOPS
Электрическая мощность:	17,808.00 кВт
Память:	1,024,000 ГБ
Внутренняя сеть:	TH Express-2
Операционная система:	Kylin Linux
Компилятор:	icc
Математическая библиотека:	Intel MKL-11.0.0
MPI:	MPICH2 (GLEX channel)



Tianhe-2 (MilkyWay-2) - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P





Titan - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x

Расположение:	Национальная лаборатория Оук-Ридж (США)
Производитель:	Cray Inc.
Количество ядер:	560,640
Производительность Linpack (Rmax)	17,590.0 teraFLOPS
Пиковая производительность (Rpeak)	27,112.5 teraFLOPS
Электрическая мощность:	8,209.00 кВт
Память:	710,144 ГБ
Внутренняя сеть:	Cray Gemini interconnect
Операционная система:	Cray Linux Environment



Titan - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x





Sequoia - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom

Расположение:	Ливерморская национальная лаборатория (США)
Производитель:	IBM
Количество ядер:	1,572,864
Производительность Linpack (Rmax)	17,173.2 teraFLOPS
Пиковая производительность (Rpeak)	20,132.7 teraFLOPS
Электрическая мощность:	7,890.00 кВт
Память:	1,572,864 ГБ
Внутренняя сеть:	Custom Interconnect
Операционная система:	Linux



Sequoia - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom





K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect

Расположение:	Институт вычислительных систем (Япония)
Производитель:	Fujitsu
Количество ядер:	705,024
Производительность Linpack (Rmax)	10,510.0 teraFLOPS
Пиковая производительность (Rpeak)	11,280.4 teraFLOPS
Электрическая мощность:	12,659.89 кВт
Память:	1,410,048 гБ
Внутренняя сеть:	Custom Interconnect
Операционная система:	Linux



K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect





Mira - BlueGene/Q, Power BQC 16C 1.60GHz, Custom

Расположение:	Аргонская национальная лаборатория (США)
Производитель:	IBM
Количество ядер:	786,432
Производительность Linpack (Rmax)	8,586.6 teraFLOPS
Пиковая производительность (Rpeak)	10,066.3 teraFLOPS
Электрическая мощность:	3,945.00 кВт
Внутренняя сеть:	Custom Interconnect
Операционная система:	Linux

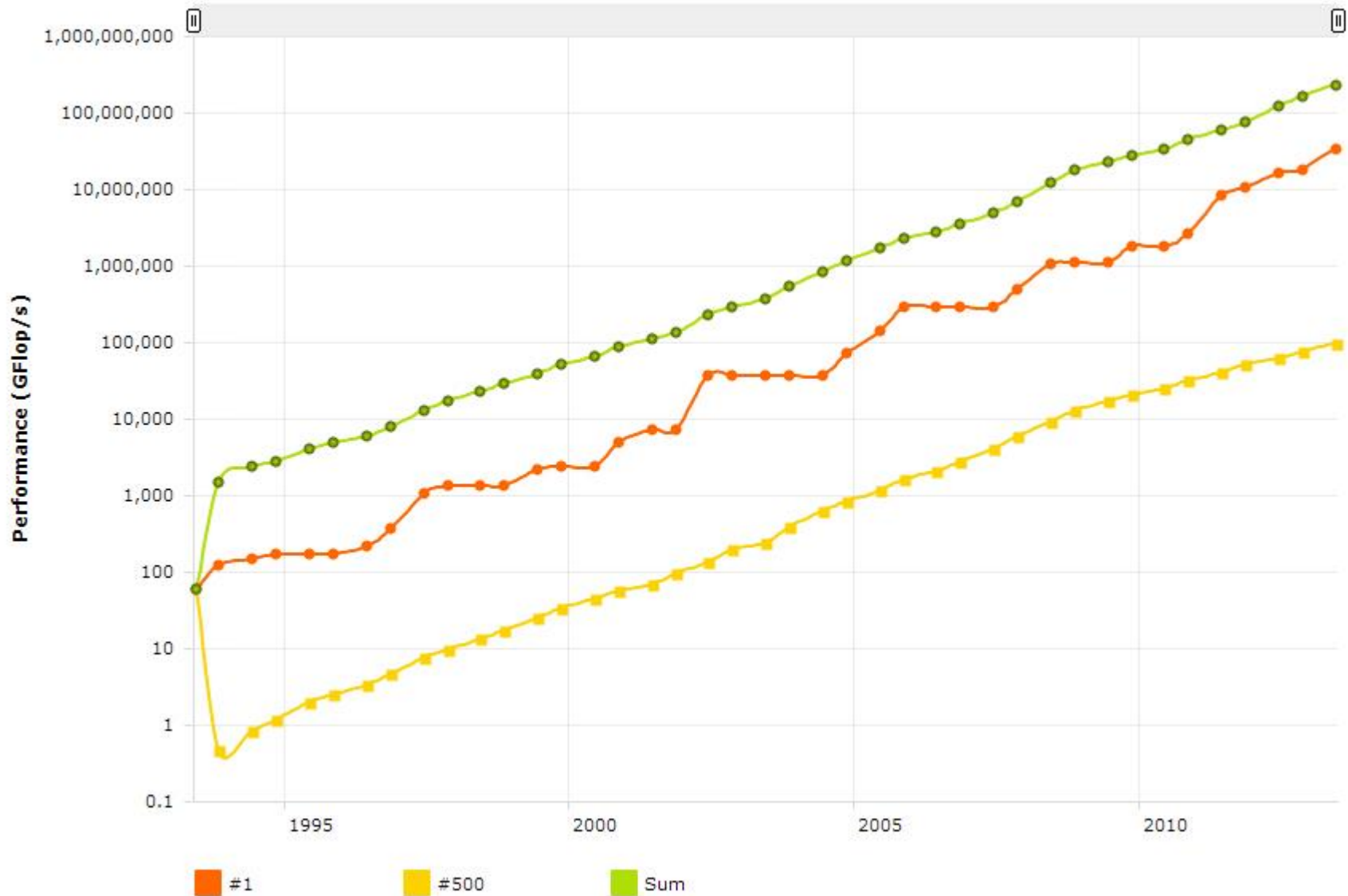


Mira - BlueGene/Q, Power BQC 16C 1.60GHz, Custom





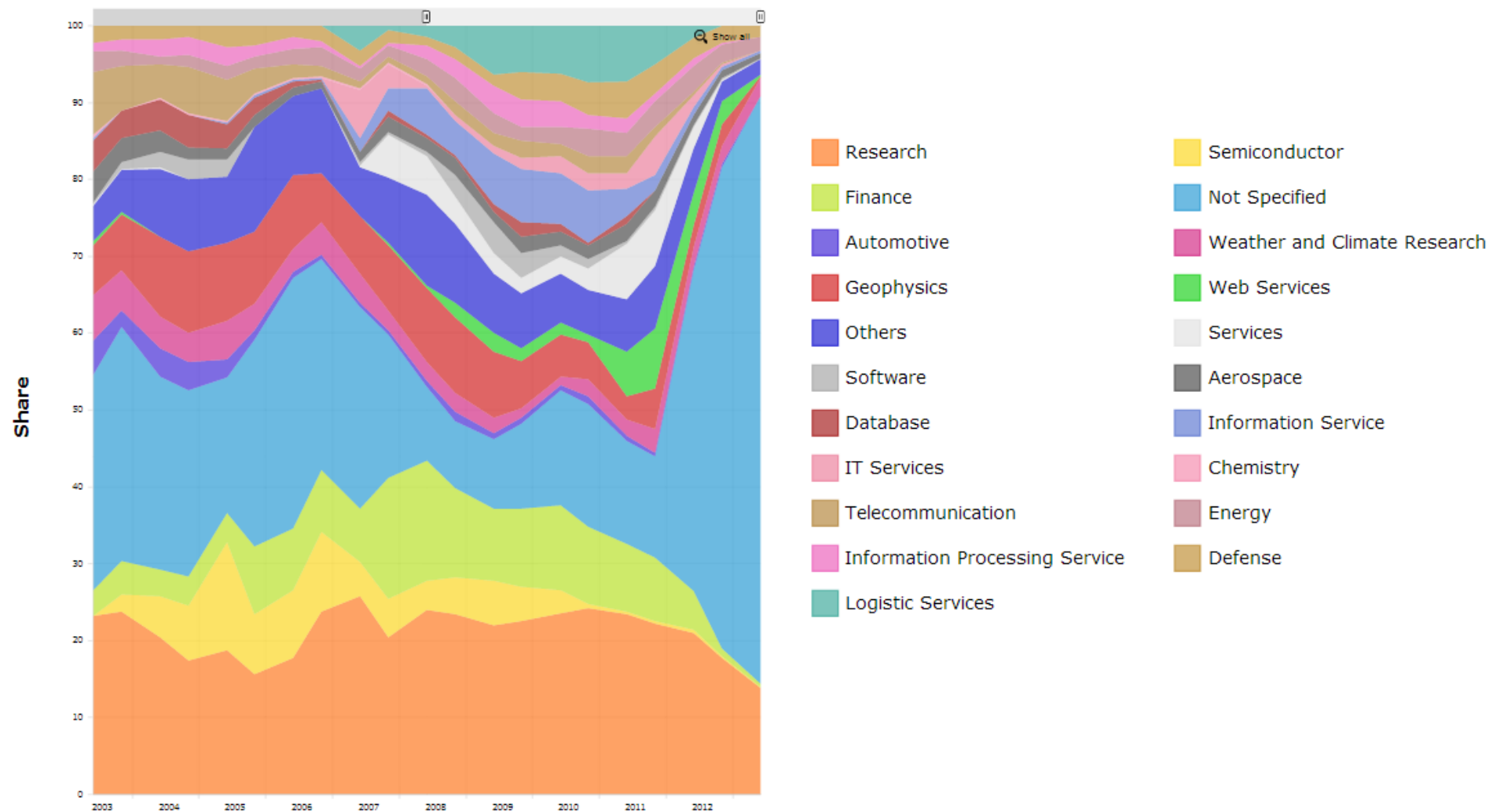
Список TOP500





Список TOP500 – области применения

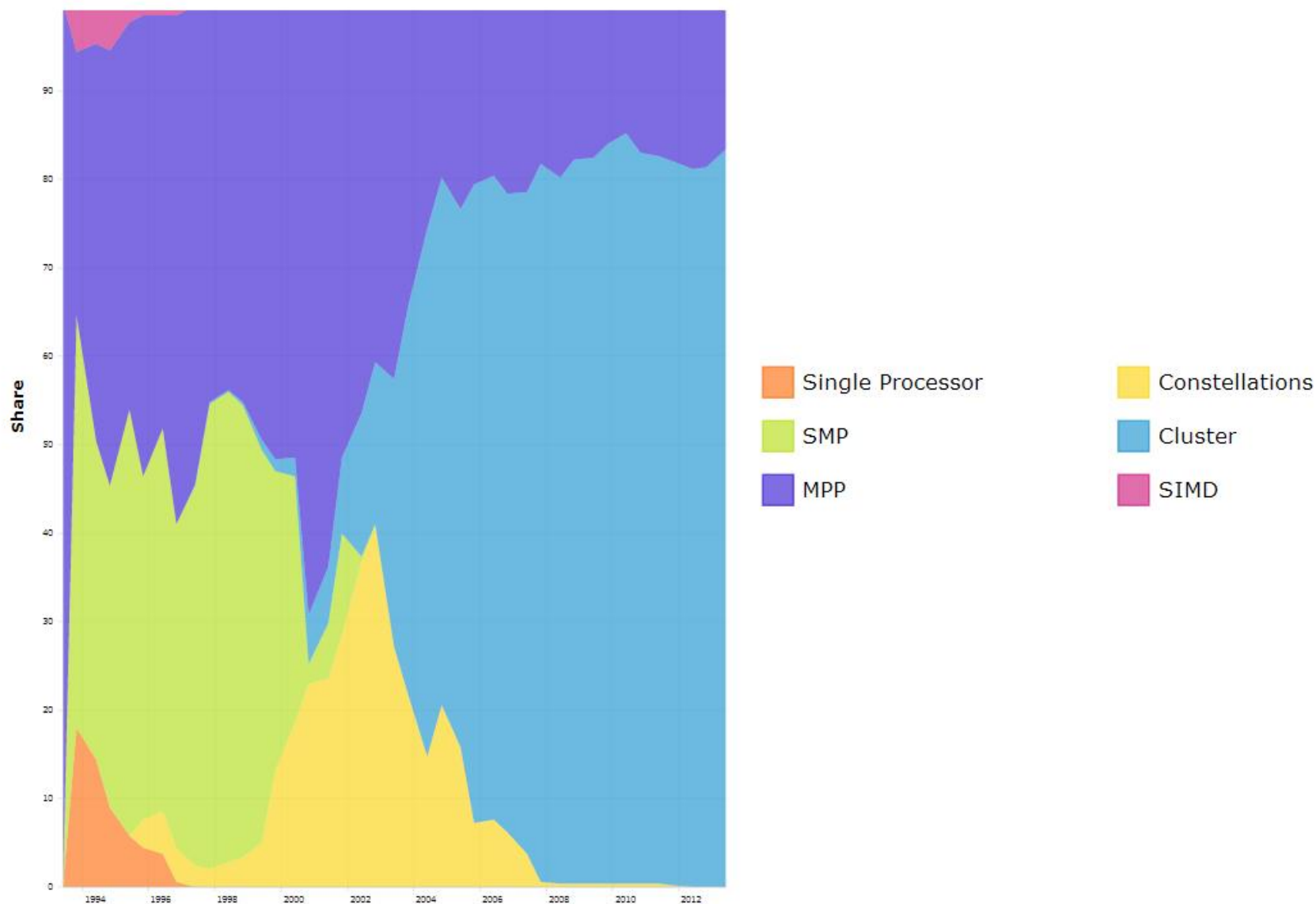
Application Area - Systems Share





Список TOP500 – типы систем

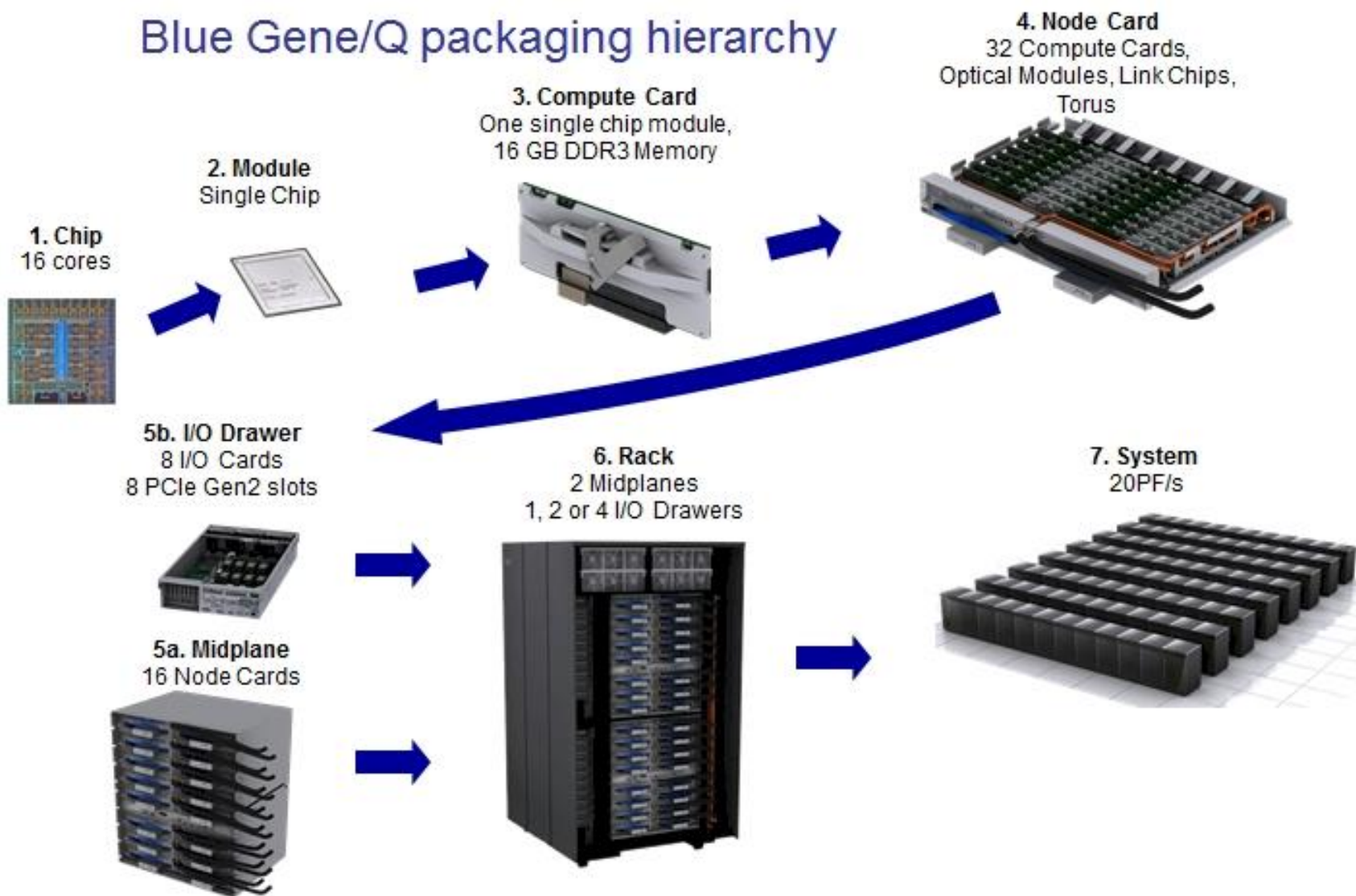
Architecture - Systems Share





Структура современных ВС

Blue Gene/Q packaging hierarchy





Структура современных ВС

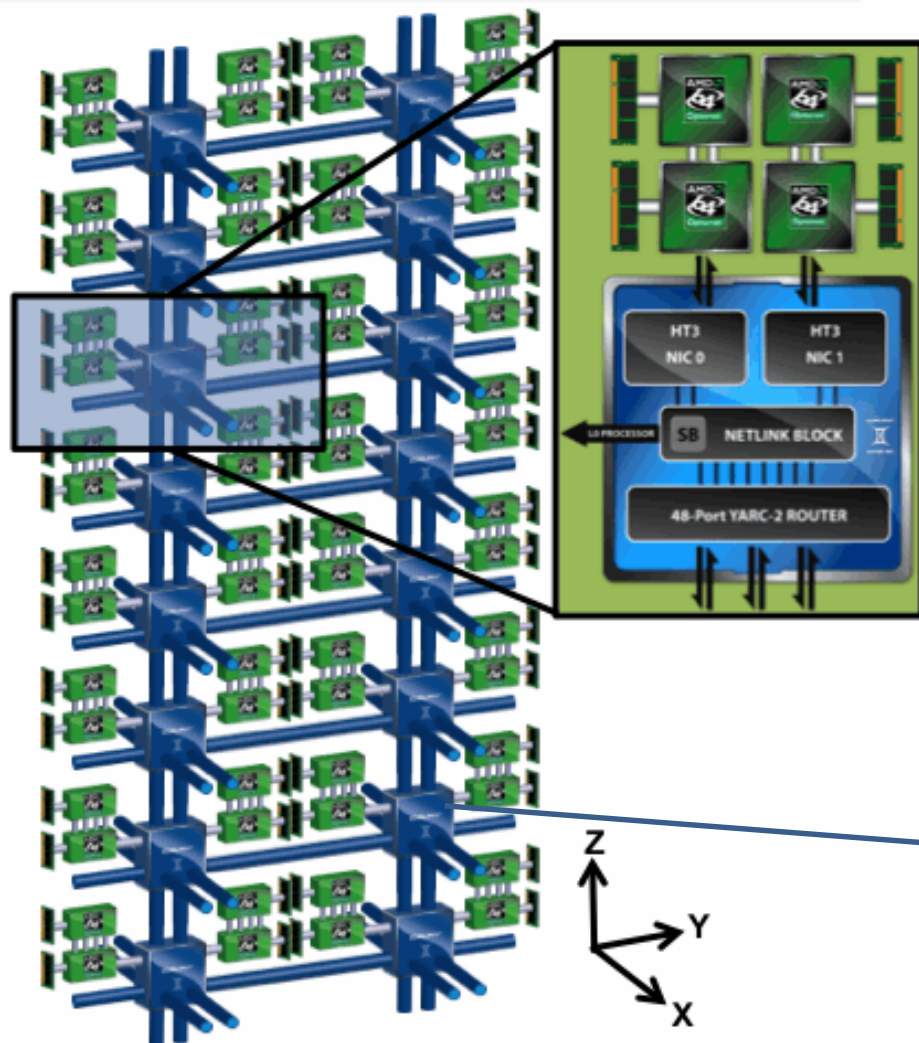
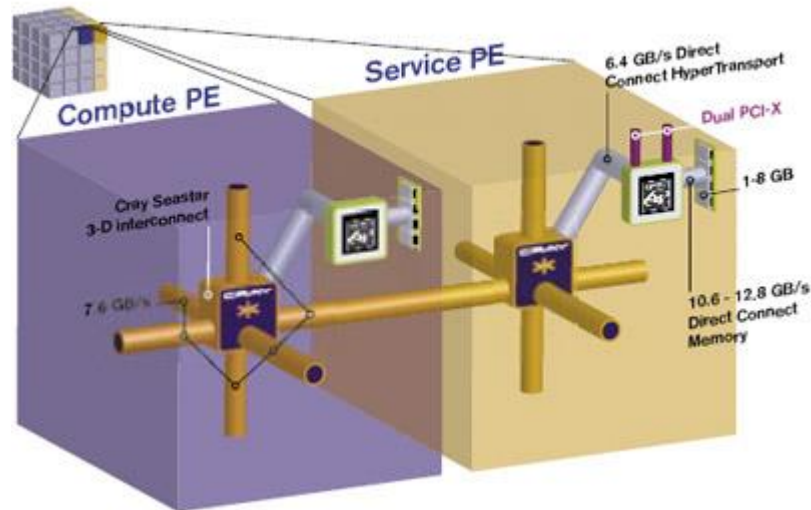


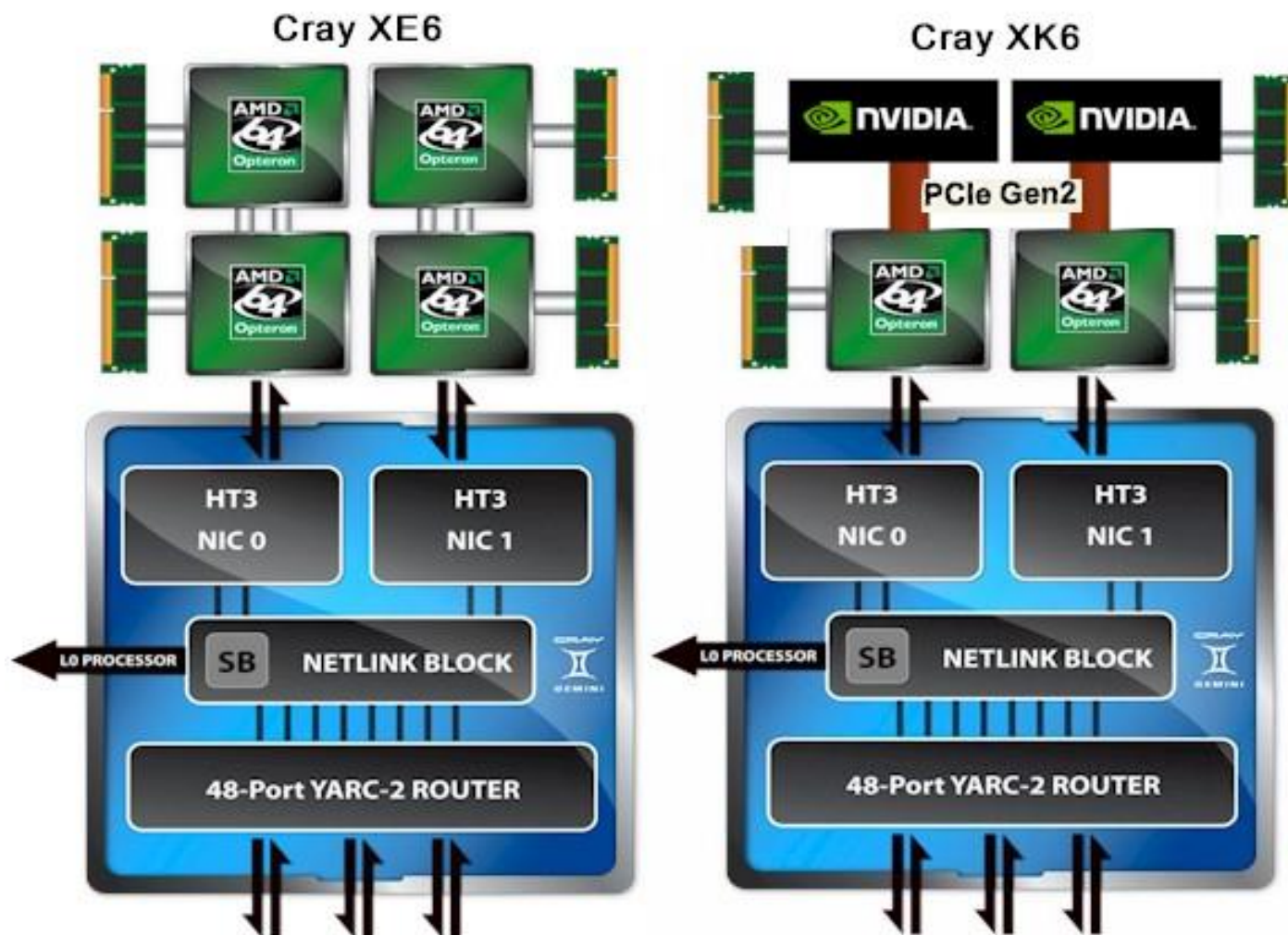
Image courtesy of Cray, Inc.

Cray XT4 Scalable Architecture





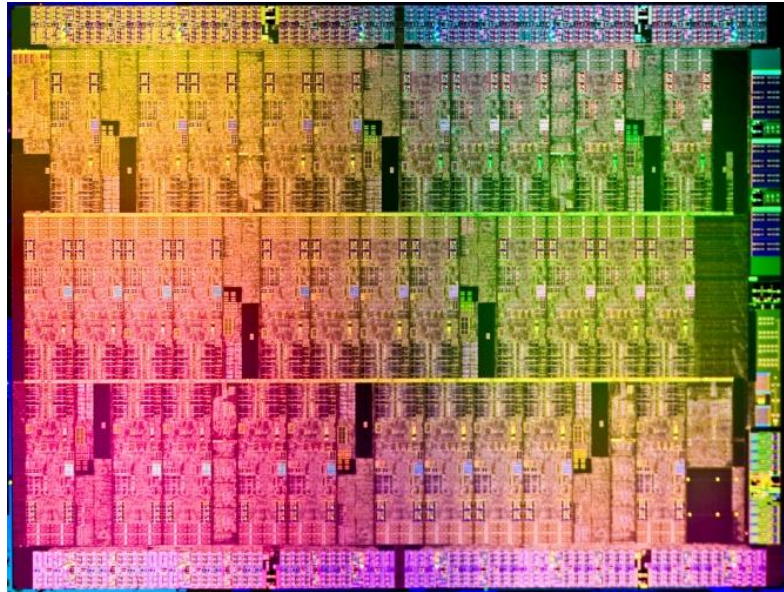
Структура современных ВС







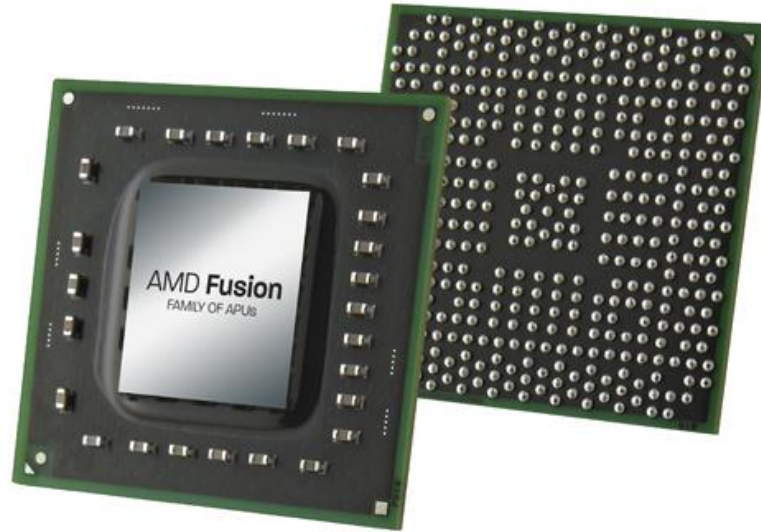
Intel MIC



- Более 50 процессорных ядер
- Более 1 teraFLOPS
- Средства разработки: OpenMP, OpenCL, Intel Cilk Plus
- 512-битные векторные АЛУ



AMD Fusion

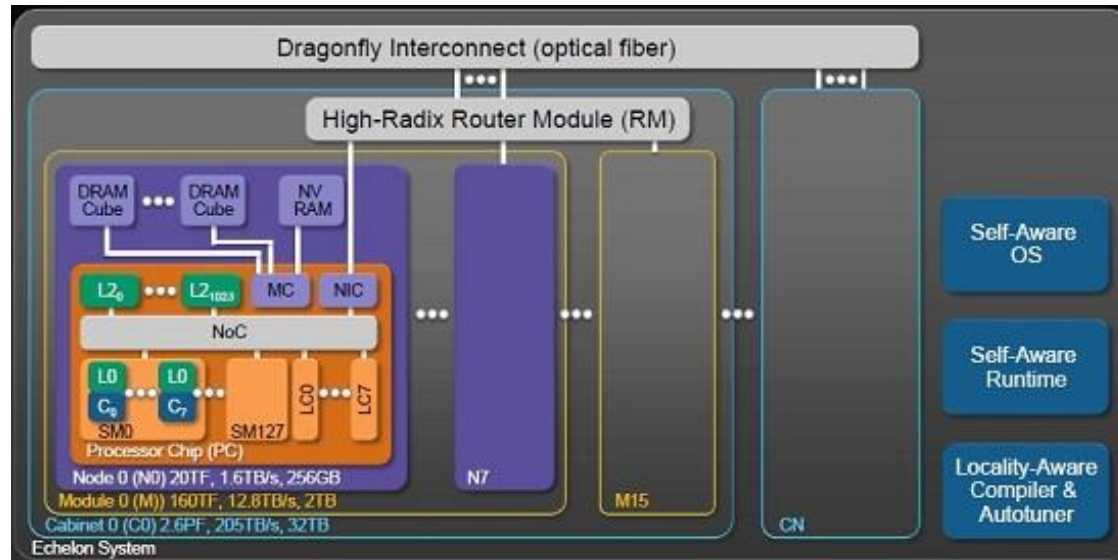


Объединение центрального многозадачного универсального процессора
с графическим параллельным многоядерным процессором
в одном кристалле

- 2-4 ядра K12
- GPU класса HD 5000
- 0,5-1 Мб кэша L2 на ядро (кэш L3 отсутствует)



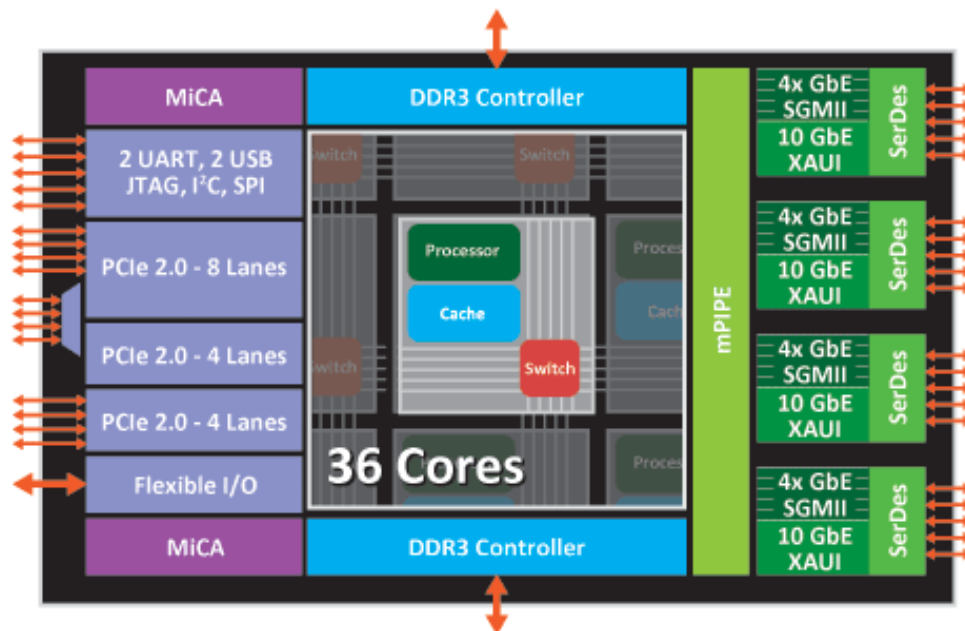
Nvidia Echelon



- 128 потоковых блоков по 8 ядер
- 4 операции двойной точности за такт
- 6-уровневый кэш объёмом 256 Мб
- Внешняя память до 256 Гб.



Tilera Tile-Gx



- До 100 ядер, связанных по не полному графу
- Трёхуровневый кэш до 26 МБ
- До 1 ТБ памяти



- **Распределённая ОС (GNU/Linux)**
- **Средства разработки параллельных программ**
(MPI, OpenMP, CUDA, Cray Chapel)
- **Системы управления ресурсами (TORQUE, SLURM)**
- **Системы обеспечения отказоустойчивости**
(самоконтроль, самодиагностика, контрольные точки)
- **Диспетчеры пространственно-распределённых ВС**
(GridWay, Pegasus)



Общая память

- POSIX Threads
- OpenMP
- Intel TBB
- Intel Cilk
- CUDA
- OpenCL

Распределённая память

- Sockets
- MPI
- PVM

PGAS

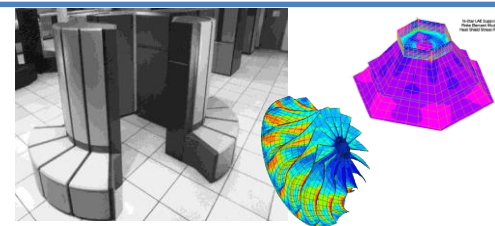
- Cray Chapel
- IBM X10
- Unified Parallel C
- HPF



Этапы развития распределённых ВС (на примере ВС фирмы Cray)

1 ГФЛОПС – 1988: Cray Y-MP; 8 процессоров

- Задачи гидродинамики



1 ТФЛОПС – 1998: Cray T3E; 1 024 процессоров

- Моделирование процессов магнетизма



1 ПФЛОПС – 2008: Cray XT5; 150 000 процессоров

- Моделирование сверхпроводимости



1 ЭФЛОПС – 2018: _____; ~10 000 000 процессоров

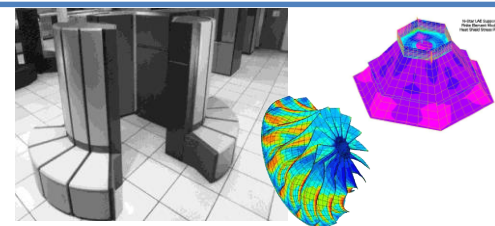
- ???



Этапы развития распределённых ВС (на примере ВС фирмы Cray)

1 ГФЛОПС – 1988: Cray Y-MP; 8 процессоров

- Задачи гидродинамики
- Fortran77 + векторизация



1 ТФЛОПС – 1998: Cray T3E; 1 024 процессоров

- Моделирование процессов магнетизма
- Fortran + MPI (Message Passing Interface)



1 ПФЛОПС – 2008: Cray XT5; 150 000 процессоров

- Моделирование сверхпроводимости
- C/C++/Fortran + MPI + векторизация



1 ЭФЛОПС – 2018: _____; ~10 000 000 процессоров

- ???
- C/C++/Fortran + MPI + CUDA/OpenCL/OpenMP/OpenACC

Или, может быть,
что-то совершенно
иное?



Цели организации функционирования:

- Минимум времени решения задачи
- Максимальная надёжность
- Максимум прибыли
- Минимизации энергопотребления
- ...

Могут быть составные показатели.



I Монопрограммный режим

Решение одной сложной задачи – для решения задачи используются все ресурсы ВС.

II Мультипрограммный режим

Обработка набора задач – учитывается не только количество задач, но их параметры: число ветвей, время решения и др.

Обслуживание потока задач – задачи поступают в случайные моменты времени, их параметры случайны.



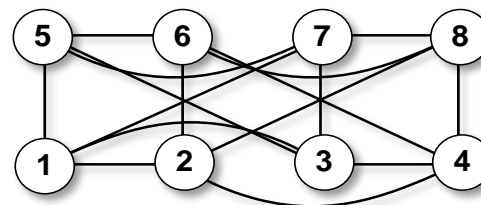
Вложение параллельных программ в ВС

Вложение High Performance Linpack в подсистему:

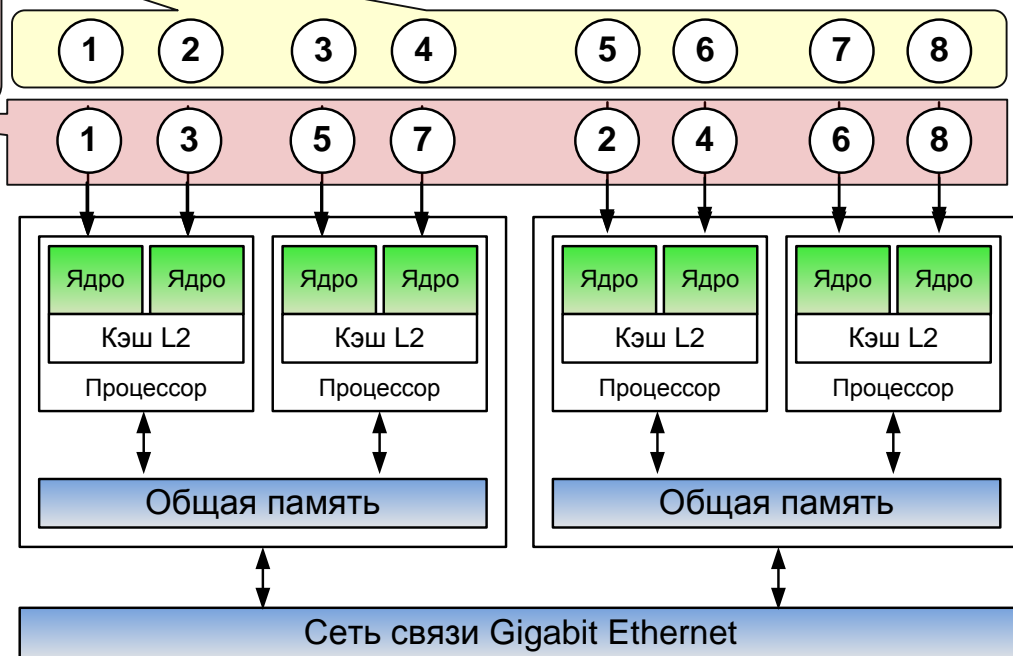
стандартными MPI-утилитами –
время выполнения 118 сек. (44 GFLOPS)

разработанными средствами –
время выполнения **100 сек. (53 GFLOPS)**

Иерархическая ВС:
2 узла по 2 Intel Xeon 5150



High Performance Linpack (HPL)





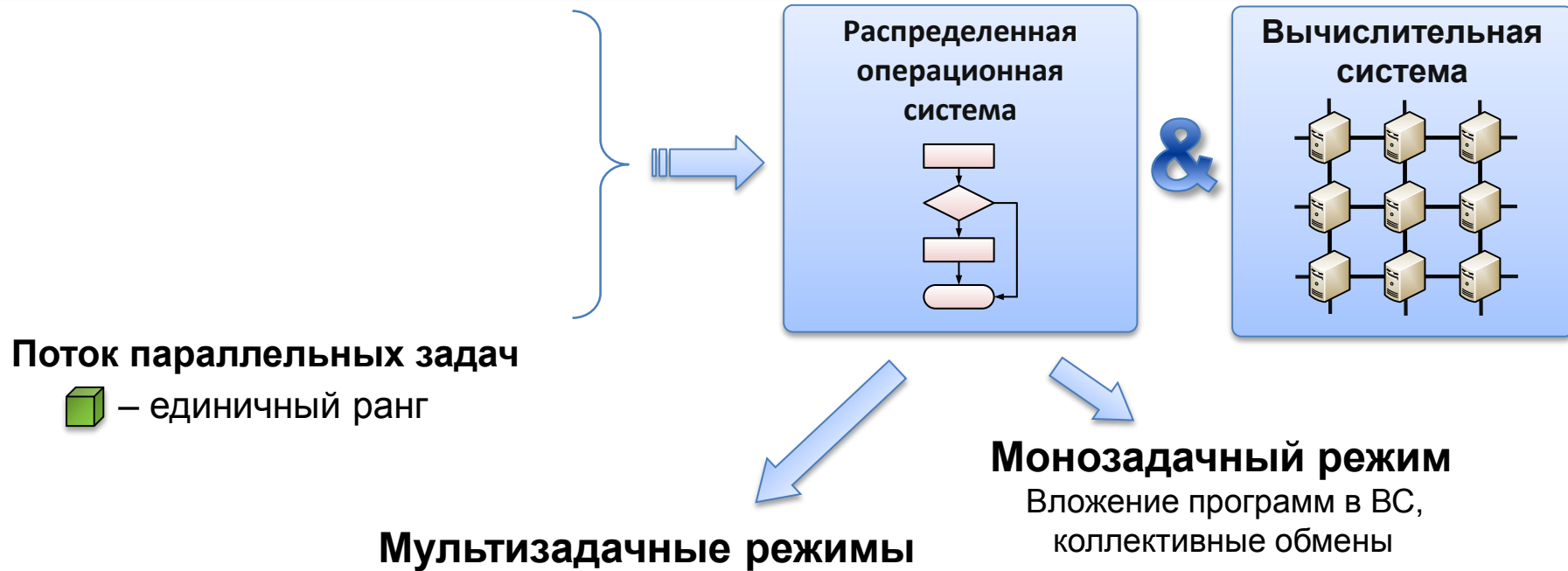
Сеанс работы пользователей в пакетном режиме

1. Поставить задачу в очередь.
2. Проверить состояние задачи.
3. Внести коррективы в задачу (её параметры).
4. Получить результат решения задачи.

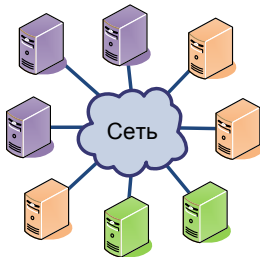
Системы управления ресурсами распределённых ВС
(RMS - Resource Management System)



Параллельное мультипрограммирование

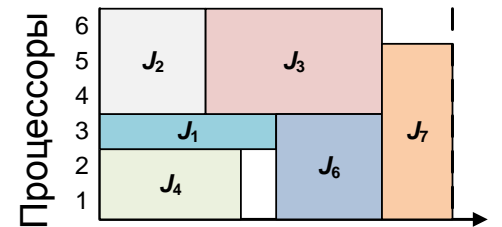


Обслуживание потоков задач
Генерация подсистем в пределах ВС



- Техника теории игр
- Стохастическое программирование

Обработка наборов задач
Формирование расписаний решения
параллельных задач



Точные, эвристические и стохастические
методы и алгоритмы



I Монопрограммный режим

Решение одной сложной задачи – для решения задачи используются все ресурсы ВС.

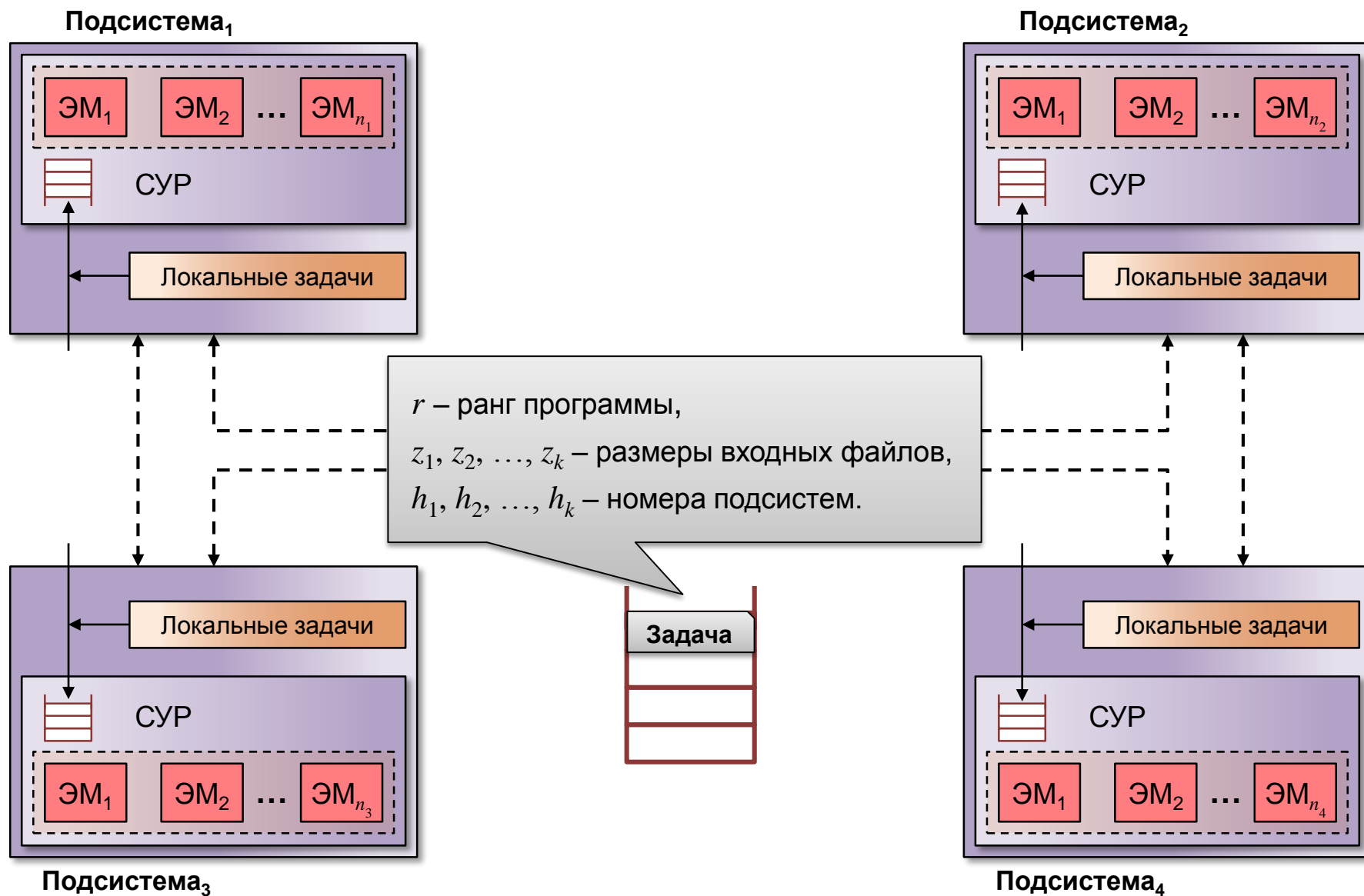
II Мультипрограммный режим

Обработка набора задач – учитывается не только количество задач, но их параметры: число ветвей, время решения и др.

Обслуживание потока задач – задачи поступают в случайные моменты времени, их параметры случайны.



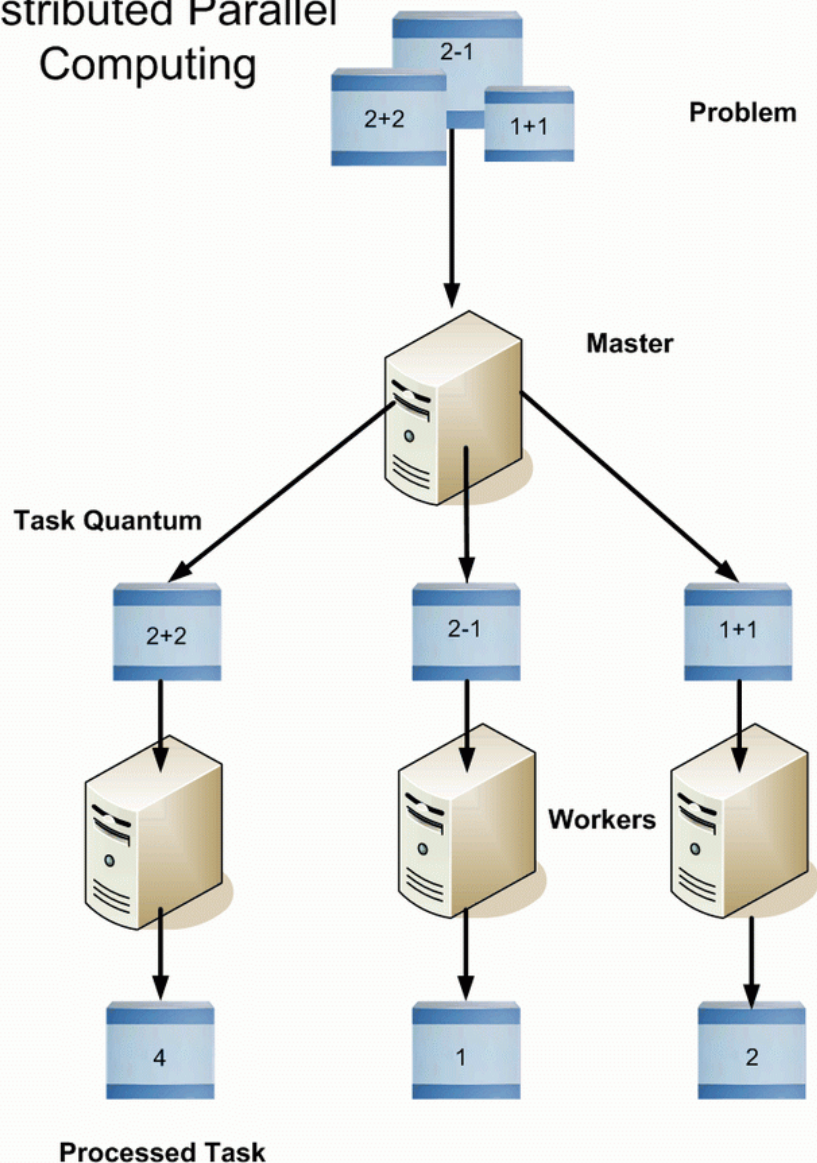
Диспетчеризация параллельных задач





Распределённые вычисления (Вычисления высокой пропускной способности, НТС)

Distributed Parallel
Computing



- живучесть, отказоустойчивость (миграция, контрольные точки)
- длительное время решения большого количества заданий
- слабо связанные задания
- большой объём задействованных ресурсов

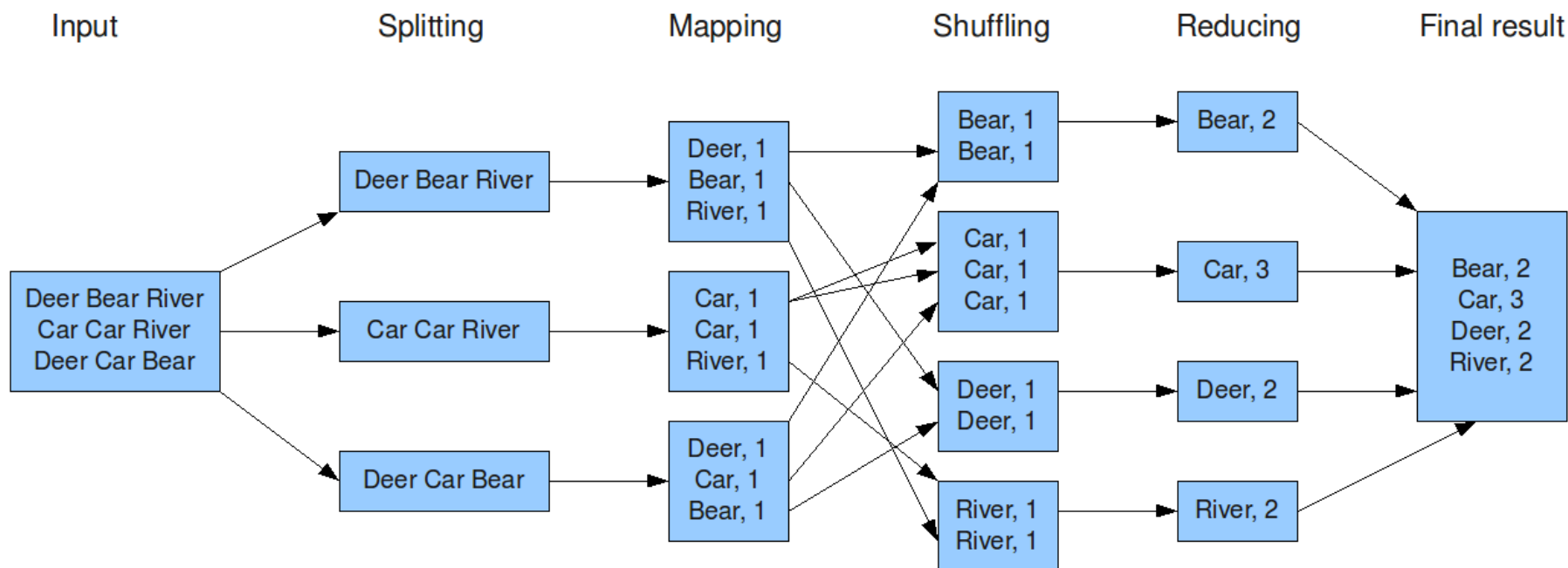
*CONDOR, MOSIX, BOINC,
MapReduce*



Распределённые вычисления (Вычисления высокой пропускной способности, HTS)

Алгоритмы MapReduce

The overall MapReduce word count process





Высокопроизводительные вычисления (HPC)



- сильно связанные параллельные ветви – интенсивный обмен сообщениями – быстрые каналы связи
- единая точка доступа к ресурсам
- параллельные программы



Конец слайдов



П.Пикассо. «Герника»