

Computer Systems Laboratory
A.V. Rzhanov Institute of
Semiconductor Physics SB RAS
Lavrentiev ave., 13
630090, Novosibirsk, Russia
Tel. & Fax: +7 (383) 333 21 71
E-mail: khor@isp.nsc.ru

Computer Center for Parallel Technologies
Siberian State University of
Telecommunications and Informatics
Kirov str., 86
630102, Novosibirsk, Russia
Tel. & Fax: +7 (383) 269 82 75
E-mail: khor@sibsutis.ru

GEOGRAPHICALLY DISTRIBUTED COMPUTER SYSTEMS AND PARALLEL MULTIPROGRAMMING

Prof. Dr. Victor KHOROSHEVSKY
Corresponding Member of
Russian Academy of Sciences

DICR'2010

Novosibirsk, 30.11.- 03.12.2010

Лаборатория вычислительных систем
Институт физики полупроводников
им. А.В. Ржанова СО РАН
пр-кт ак. Лаврентьева, 13
630090, Новосибирск, Россия
Тел. & факс: +7 (383) 333 21 71
E-mail: khora@isp.nsc.ru

Центр параллельных
вычислительных технологий
Сибирский государственный университет
телекоммуникаций и информатики
ул. Кирова, 86
630102, Новосибирск, Россия
Тел. & факс: +7 (383) 269 82 75
E-mail: khora@sibsutis.ru

В.Г. Хорошевский
член-корреспондент Российской академии наук

**ПРОСТРАНСТВЕННО-РАСПРЕДЕЛЕННЫЕ
ВЫЧИСЛИТЕЛЬНЫЕ СИСТЕМЫ И ПАРАЛЛЕЛЬНОЕ
МУЛЬТИПРОГРАММИРОВАНИЕ**

**Распределенные информационные и вычислительные
ресурсы**

декабрь 10

Новосибирск, 30 ноября – 3 декабря 2010 г.

CONFERENCE TOPICS

1. Multi-architecture of high-performance computer systems (CS)

2. Geographically distributed multicluster computer system

3. Parallel multiprogramming

- **Parallel program mapping**
- **Moldable jobs scheduling**
- **Game-theoretic model**
- **Stochastic programming**

МУЛЬТИАРХИТЕКТУРА СОВРЕМЕННЫХ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ

№	Название системы	Производительность, GFLOPS	Количество ядер	Вычислительный узел	Тип системы	Структура сети
1	Jaguar Cray XT5-HE	2 331 000	224162	2 x AMD Opteron six-core	MPP	3D-тор
2	Nebulae Dawning TC3600 Blade	2 984 300	120640	2 x Intel Xeon X56xx, Nvidia Tesla C2050	Кластер	Двухуровневая (fat tree)
3	RoadRunner IBM BladeCenter QS22/LS21	1 375 780	122400	2 x AMD Opteron dual core, 4 x IBM PowerXCell 8i	Мультикластер (18 кластеров)	Двухуровневая (fat tree)
4	Kraken XT5 Cray XT5-HE	1 028 850	98928	2 x AMD Opteron six-core	MPP	3D-тор
5	JUGENE IBM BlueGene/P	1 002 700	294912	4 x IBM PowerPC 450	MPP	3D-тор, бинарное дерево

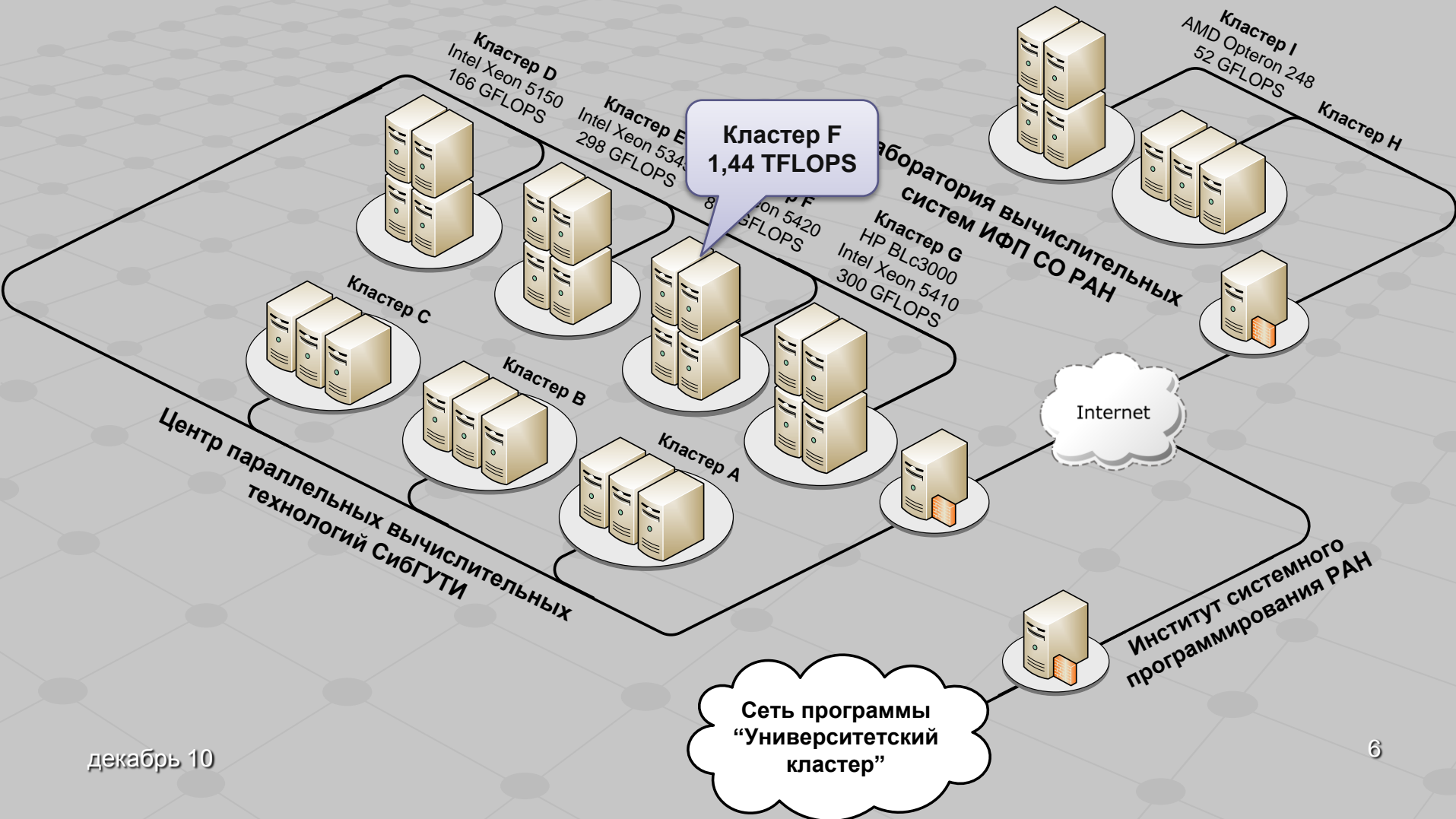
МУЛЬТИАРХИТЕКТУРА СОВРЕМЕННЫХ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ

№	Название системы	Производительность, GFLOPS	Количество ядер	Вычислительный узел	Тип системы	Структура сети
1	Tianhe-1A NUDT YH MPP	4 701 000	186 368	2 x Intel Xeon X5670, NVidia M2050	Кластер	Двухуровневая (fat tree)
2	Jaguar Cray XT5-HE	2 331 000	224 162	2 x AMD Opteron six-core	MPP	3D-тор
3	Nebulae Dawning TC3600 Blade	2 984 300	120 640	2 x Intel Xeon X56xx, NVidia Tesla C2050	Кластер	Двухуровневая (fat tree)
4	TSUBAME 2 HP ProLiant SL390s G7	2 287 630	73 278	2 x Intel Xeon X56xx, NVidia Tesla M2050/S1070	Кластер	Двухуровневая (fat tree)
5	Hopper Cray XE6	1 288 630	153 408	2 x AMD Opteron 12 core	MPP	3D-тор

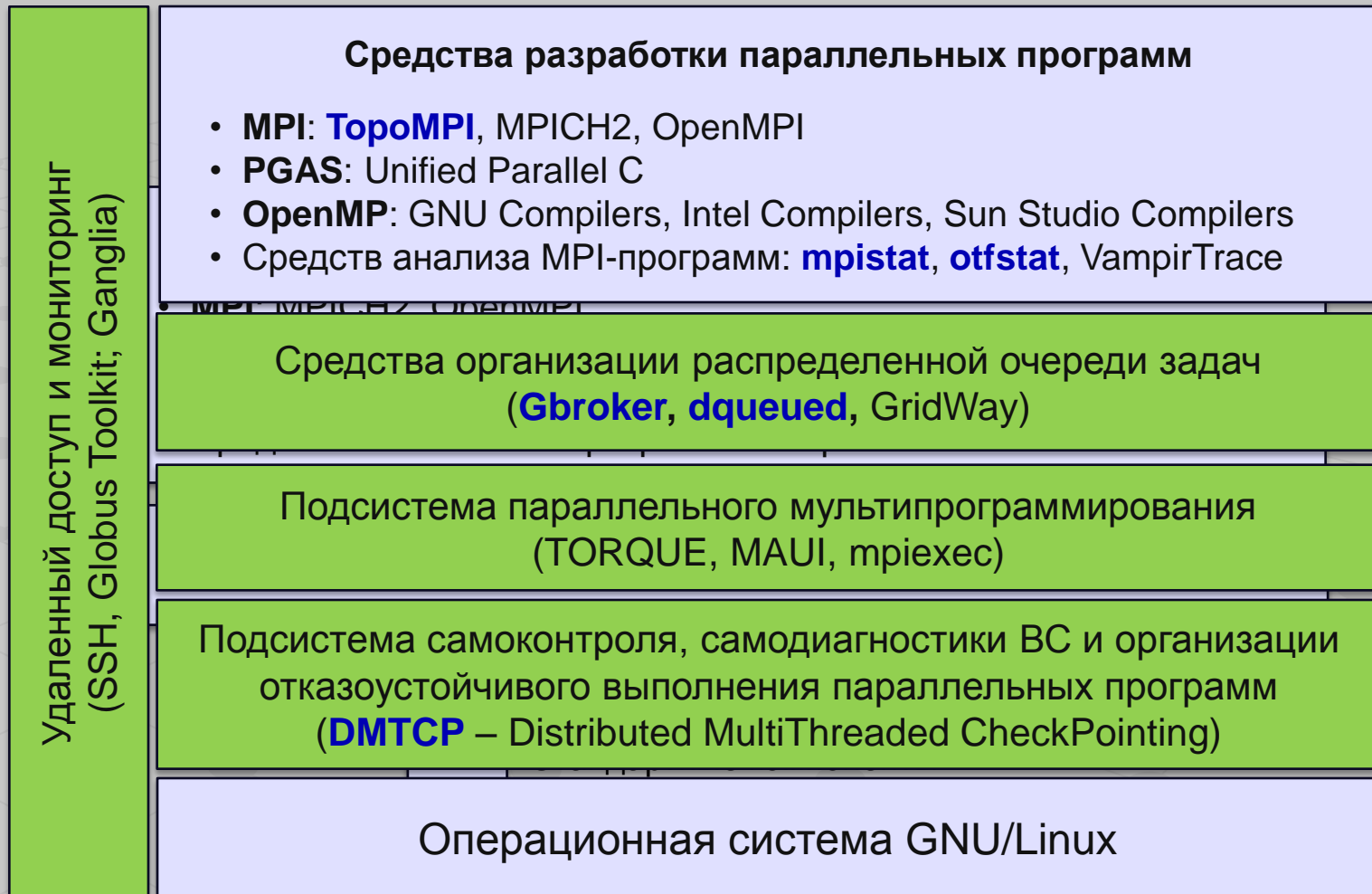
Пространственно-распределённая мультикластерная вычислительная система *GRID-модель*

Парадигмы:

- Программируемость структуры, масштабируемость, живучесть
- Параллельное мультипрограммирование



ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ МУЛЬТИКЛАСТЕРНОЙ ВС



Подсистема параллельного мультипрограммирования



Разрабатываемые в ЦПВТ ГОУ ВПО “СибГУТИ” компоненты

ПАРАЛЛЕЛЬНОЕ МУЛЬТИПРОГРАММИРОВАНИЕ

Режимы функционирования ВС


- *Монопрограммный режим*
Решение одной сложной задачи, представленной параллельной программой
Крупноблочное распараллеливание
- *Мультипрограммные режимы*
 - Обработка набора задач
 - Обслуживание потока задач

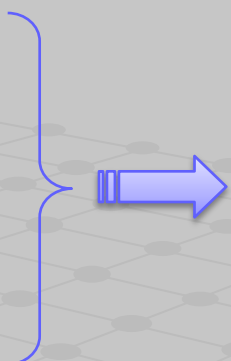
Первые работы

- *В.Г. Хорошевский.* Об алгоритмах распределения задач по ЭЦВМ // Труды СФТИ. Томск: ТГУ, 1965. Вып. 47
- *Д.А. Поспелов.* Теоретические проблемы, связанные с объединением типовых вычислительных машин в единую систему // Вычислительные системы, труды симпозиума. Новосибирск: ИМ СО АН СССР. 1967.

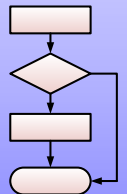
ПАРАЛЛЕЛЬНОЕ МУЛЬТИПРОГРАММИРОВАНИЕ

Поток параллельных задач

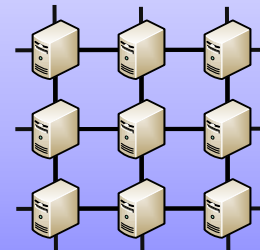
 – единичный ранг



Распределенная операционная система



Вычислительная система

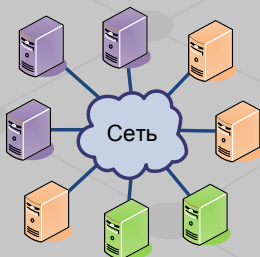


Мультипрограммные режимы

Монопрограммный режим

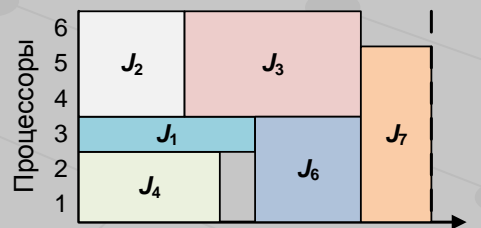
Обслуживание потоков задач

Генерация подсистем в пределах ВС



Обработка наборов задач

Формирование расписаний решения параллельных задач



Точные, эвристические и стохастические методы и алгоритмы

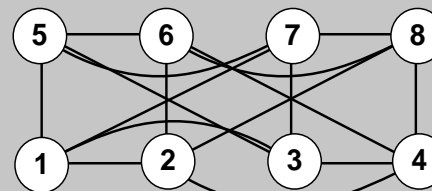
- Техника теории игр
- Стохастическое программирование

ВЛОЖЕНИЕ ПАРАЛЛЕЛЬНЫХ ПРОГРАММ В ВС

Вложение High Performance Linpack в подсистему:

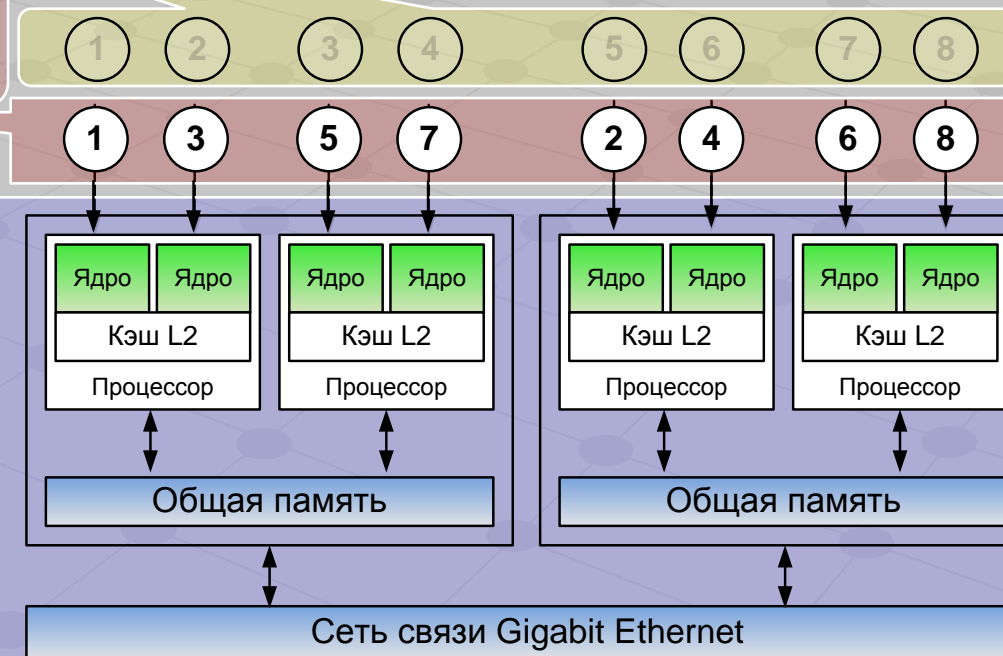
стандартными MPI-утилитами –
время выполнения **118 сек. (44 GFLOPS)**

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



Граф программы

High Performance Linpack

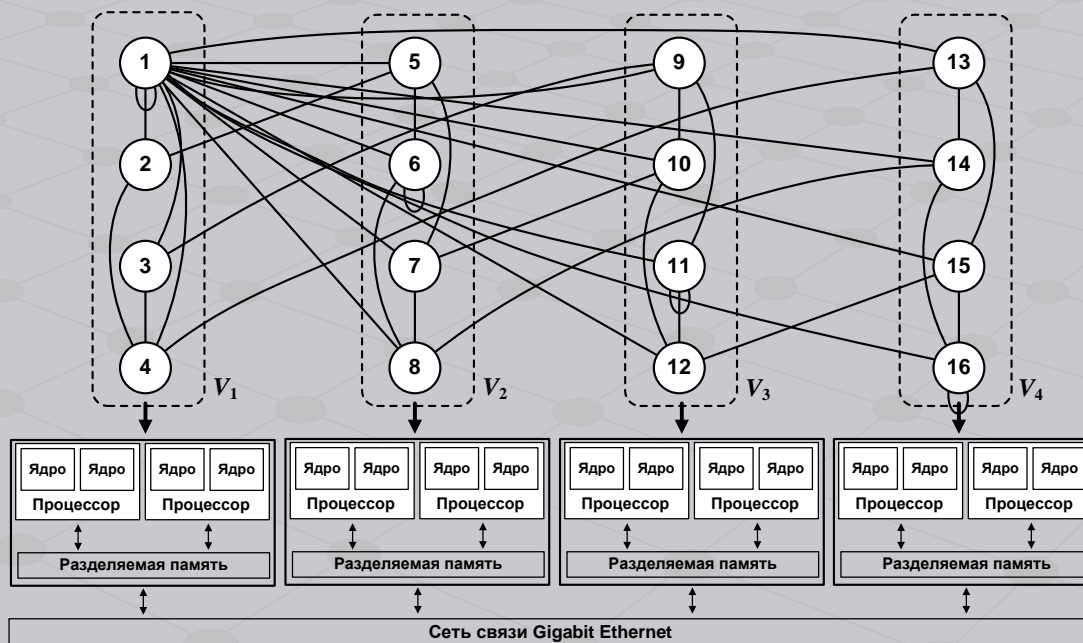


ВЛОЖЕНИЕ ПАРАЛЛЕЛЬНЫХ ПРОГРАММ В ВС

Метод вложения основан на многоуровневых (multilevel) алгоритмах разбиения информационных графов $G = (V, E)$ параллельных программ.

1. Граф G разбивается на k подмножеств; k – отношение числа L ветвей программы к числу q ядер, составляющих узел ВС. В каждое из подмножеств включаются ветви, обменивающиеся **большими объемами данных**.

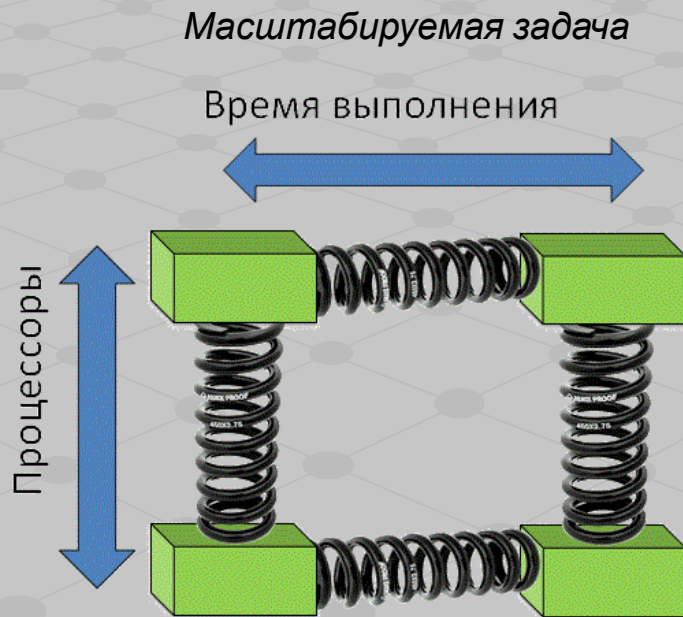
2. Параллельные ветви из i -го подмножества распределяются по ядрам i -го вычислительного узла, $i \in \{1, 2, \dots, k\}$.



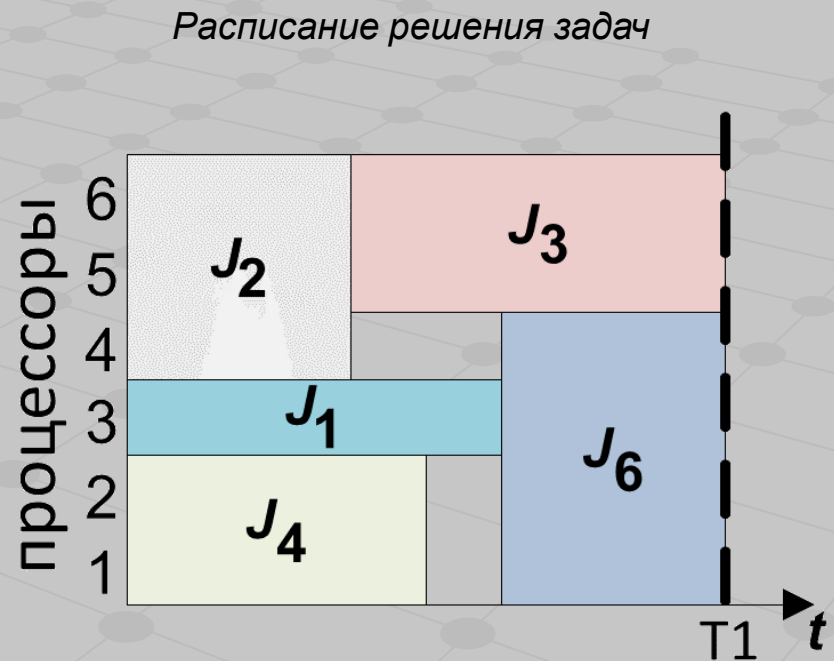
Вложение MPI-программы Conjugate Gradient из пакета NAS Parallel Benchmarks, реализующей решение системы линейных алгебраических уравнений методом сопряженных градиентов в вычислительный кластер: $L = 16$; $q = 4$; $k = 4$

ФОРМИРОВАНИЕ РАСПИСАНИЙ РЕШЕНИЯ МАСШТАБИРУЕМЫХ ЗАДАЧ

Алгоритмы основаны на методе разбиения набора масштабируемых задач на пакеты и учитывают предпочтение пользователей по выбору значений параметров задач (ранг и время решения).



Свойством масштабируемости обладают **более 80% задач**, решаемых на вычислительных системах.



$(T_2 - T_1)$ – выигрыш по времени решения задач набора

ОРГАНИЗАЦИЯ СТОХАСТИЧЕСКИ ОПТИМАЛЬНОГО ФУНКЦИОНИРОВАНИЯ ВС

Теоретико-игровой подход

1. Постановка проблемы. Поток задач с очередью.

Имеется:

- распределенная ВС из N ЭМ;
- очереди задач всех рангов;
- операционная система (ОС) – для распределения задач по машина ВС.

Задача ранга j требует для своего решения подсистему из j ЭМ.

2. Простейшая игровая модель. Игра двух объектов: ВС & ОС

i – чистая стратегия ВС; i машин выделяется для решения задач

j – чистая стратегия ОС; выбирается задача ранга j для решения на ВС

$$C = \left\| c_{ij} \right\| \quad - \text{ матрица платежей, } i, j \in \{0, 1, \dots, N\}$$

c_{ij} – платеж при выборе стратегий i и j соответственно ВС и ОС

$$\pi = \{ \pi_0, \pi_1, \dots, \pi_N \} \quad - \text{ смешанная стратегия ВС}$$

$$p = \{ p_0, p_1, \dots, p_N \} \quad - \text{ смешанная стратегия ОС}$$

ОРГАНИЗАЦИЯ СТОХАСТИЧЕСКИ ОПТИМАЛЬНОГО ФУНКЦИОНИРОВАНИЯ ВС

Теоретико-игровой подход (продолжение)

3. Оптимальные смешанные стратегии

Средний платеж ВС –

$$\sum_{i=0}^N \sum_{j=0}^N c_{ij} p_i \pi_j = p^T C \pi,$$

если ВС и ОС используют смешанные стратегии p и π соответственно.

Существуют оптимальные смешанные стратегии p^* и π^* такие, что

$$p^T C \pi^* \leq v \text{ для всех } p, \quad (p^*)^T C \pi \geq v \text{ для всех } \pi$$

Цена игры $v = (p^*)^T C \pi^*$

Элементы матрицы C :

$$c_{ij} = \begin{cases} jc_1 + (i-j)c_2 & \text{для } i \geq j, \\ ic_2 + (j-i)c_3 & \text{для } i < j, \end{cases}$$

c_1 - платеж за использование одной ЭМ в единицу времени,
 c_2 и c_3 - штрафы в единицу времени за простой одной ЭМ и при $j - i = 1$

Теорема. Матрица C не имеет седловых точек тогда и только тогда, когда

$$c_1 < \min \{c_2, c_3\}.$$

4. Параллельный алгоритм решения проблемы основывается на композиции симплекс-метода и модифицированного метода Брауна-Робинсон.

ОРГАНИЗАЦИЯ СТОХАСТИЧЕСКИ ОПТИМАЛЬНОГО ФУНКЦИОНИРОВАНИЯ РАСПРЕДЕЛЕННЫХ ВС

Техника стохастического программирования

1. Проблема организации подсистем

N - число ЭМ, образующих ВС

L - число терминалов, воспринимающих поток задач

a_{jl} - число подсистем ранга j , которое требуется терминалу l

$p_{jl}(a)$ - плотность распределения вероятностей случайной величины a_{jl} ,

$$\int_0^{\infty} p_{jl}(a) da = 1, \quad j \in \{1, 2, \dots, N\}, \quad l \in \{1, 2, \dots, L\}$$

d_{jl} - цена эксплуатации подсистемы ранга j для терминала l

c_{jl} - стоимость формирования и обслуживания подсистемы ранга j для терминала l

y_{jl} - число подсистем ранга j , обязательно выделяемых терминалу l

x_{jl} - число подсистем ранга j , дополнительно выделяемых терминалу l

ОРГАНИЗАЦИЯ СТОХАСТИЧЕСКИ ОПТИМАЛЬНОГО ФУНКЦИОНИРОВАНИЯ РАСПРЕДЕЛЕННЫХ ВС

Техника стохастического программирования (продолжение)

Ожидаемая прибыль от эксплуатации подсистем ранга j с терминала l :

$$r_{jl}(x_{jl}) = (d_{jl} - c_{jl})(x_{jl} + y_{jl}) - d_{jl} \int_0^{x_{jl} + y_{jl}} (x_{jl} + y_{jl} - a) p_{jl}(a) da$$

Проблема:

$$\sum_{j=1}^n \sum_{l=1}^L r_{jl}(x_{jl}) \rightarrow \max_{\{x_{jl}\}}; \quad j = \overline{1, n}; \quad l = \overline{1, L};$$

$$\sum_{j=1}^n \sum_{l=1}^L jx_{jl} \leq n, \quad n = N - \sum_{j=1}^N \sum_{l=1}^L jy_{jl},$$

где $x_{jl} = 0$ для $j = \overline{n+1, N}$

2. Параллельный алгоритм решения проблемы основывается на технике динамического программирования

ЗАКЛЮЧЕНИЕ

- Распределенная вычислительная система – большемасштабный вероятностный объект, обслуживающий стохастические потоки параллельных задач
- Техника теории игр и стохастическое программирование составляют основу для организации стохастически оптимального использования ресурсов ВС
- Стохастическая оптимизация функционирования распределенных ВС осуществляется однократно для достаточно большого интервала времени
- Параллельные алгоритмы и теоретико-игровые, и стохастического программирования реализуются эффективно на распределенных ВС
- Нет сложных вычислительных проблем при создании распределенной операционной системы, поддерживающей параллельное мультипрограммирование
- Разработанный алгоритмический и программный инструментарий вложения параллельных программ в мультиархитектурные ВС эффективнее стандартных MPI-утилит

THANK YOU VERY MUCH

NOVOSIBIRSK, 30.11.- 03.12.2010
