Fortan OpenMP/DVM - язык параллельного программирования для кластеров

В.А. Бахтин, Н.А. Коновалов, В.А. Крюков, Н. В. Поддерюгина

Институт прикладной математики им. М.В.Келдыша РАН

e-mail: dvm@keldysh.ru
http://www.keldysh.ru/pages/dvm

OpenMP Fortran

- Высокоуровневая модель параллелизма с общей памятью
- Директивы, функции системы поддержки, системные переменные
- Спецкомментарии

Недостатки:

- Локализация данных и вычислений
- Явная синхронизация общих данных

Fortran-DVM

- Высокоуровневая модель параллелизма без явной ориентации на общую или распределенную память
- Директивы - спецкомментарии
- Согласованное распределение данных и вычислений (локализация)
- Не требует явной синхронизации при работе с общими данными

Fortran OpenMP/DVM

Цели:

- Расширение сферы применения модели DVM (OpenMP-программы)
- Расширение сферы использования OpenMP (системы с распределенной памятью)

Директивы распределения данных и вычислений

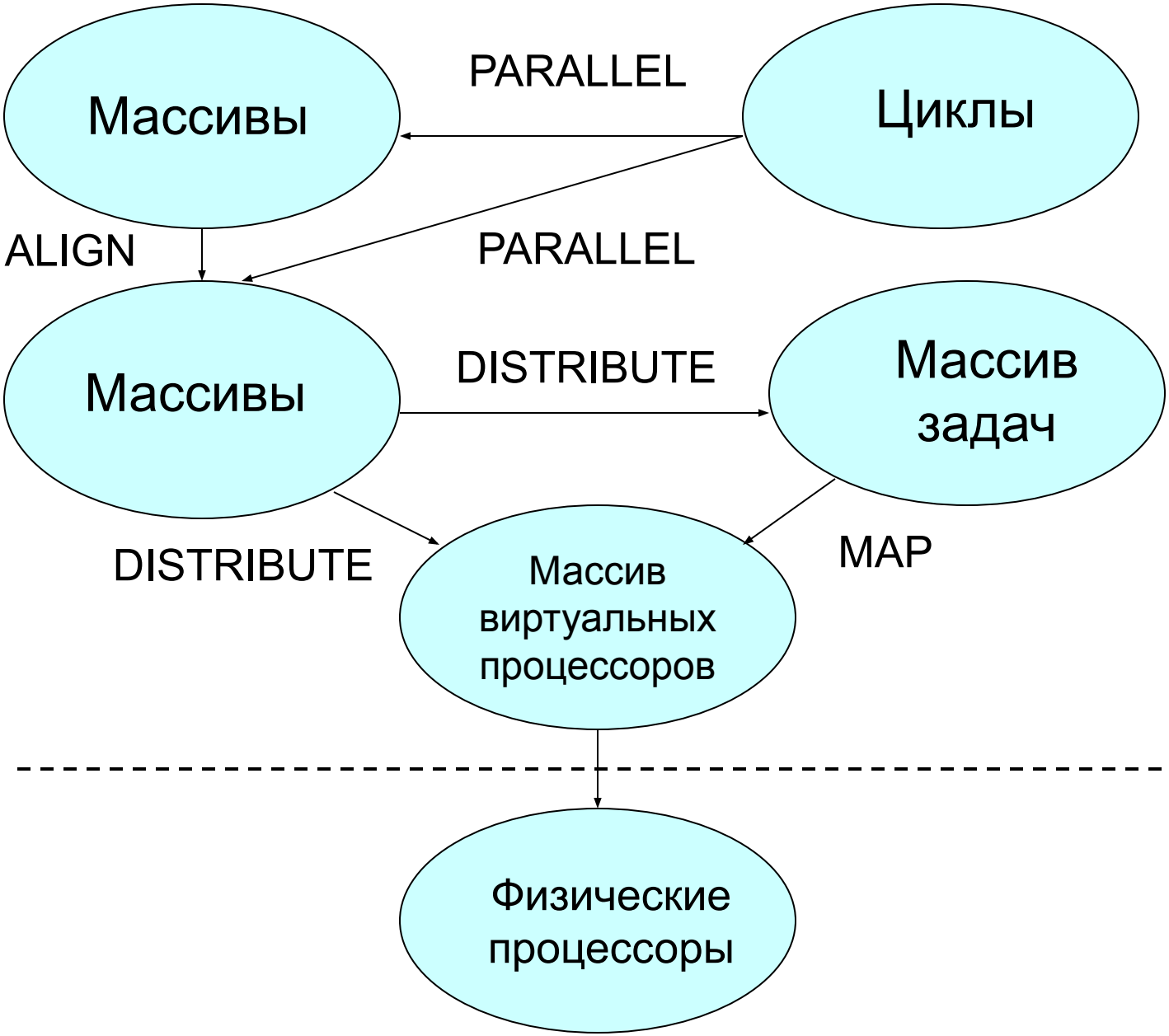
DISTRIBUTE - распределение массива на многомерную решетку виртуальных процессоров

ALIGN - распределение массива в соответствии с распределением другого массива

PARALLEL - распределение витков цикла в соответствии с распределением массива

MAP - распределение задач на секции решетки виртуальных процессоров

Отображение последовательной программы



Общие данные

REDUCTION - редукционные данные

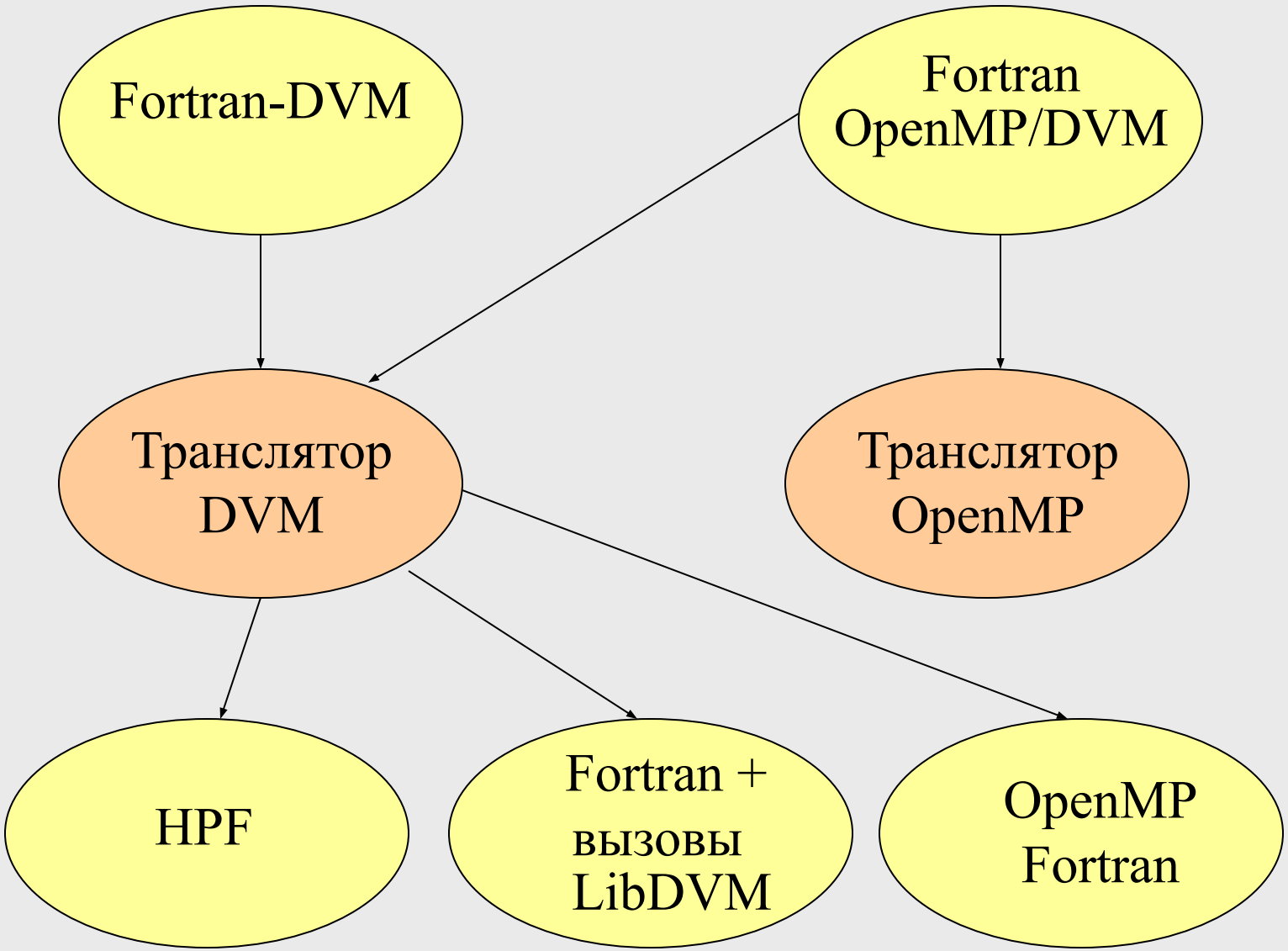
CONSISTENT - консистентные данные

SHADOW - «соседние» данные

ACROSS - «соседние» данные
с информационными связями

REMOTE - удаленные данные

Схема компиляции



Распределение данных

Рассмотрим некоторую дискретную область моделирования (массив). Если в каждой точке модели выполняется одинаковое количество вычислений, то мы будем называть эти вычисления *однородными*, иначе *неоднородными*.

real A(12), B(6)

Распределение массивов с однородными вычислениями описывается директивой DISTRIBUTE:

```
CDVM$   DISTRIBUTE  A(BLOCK)
CDVM$   DISTRIBUTE  B(BLOCK)
```

	node1	node2	node3	node4
A	1,2,3	4,5,6	7,8,9	10,11,12
B	1,2	3,4	5	6

Распределение данных

```
real B(6), WB(6)
```

Распределение массива с неоднородными вычислениями описывается директивой:

```
CDVM$  DISTRIBUTE  B(WGT_BLOCK(WB,6))
       data  WB /1., 0.5, 0.5, 0,5, 0.5, 1./
```

	node1	node2	node3	node4
B	1	2,3	4,5	6

Данные и вычисления распределяются таким образом, чтобы суммы весов вычислений на каждом процессоре были пропорциональны весам (производительности) процессоров.

Тесты NAS

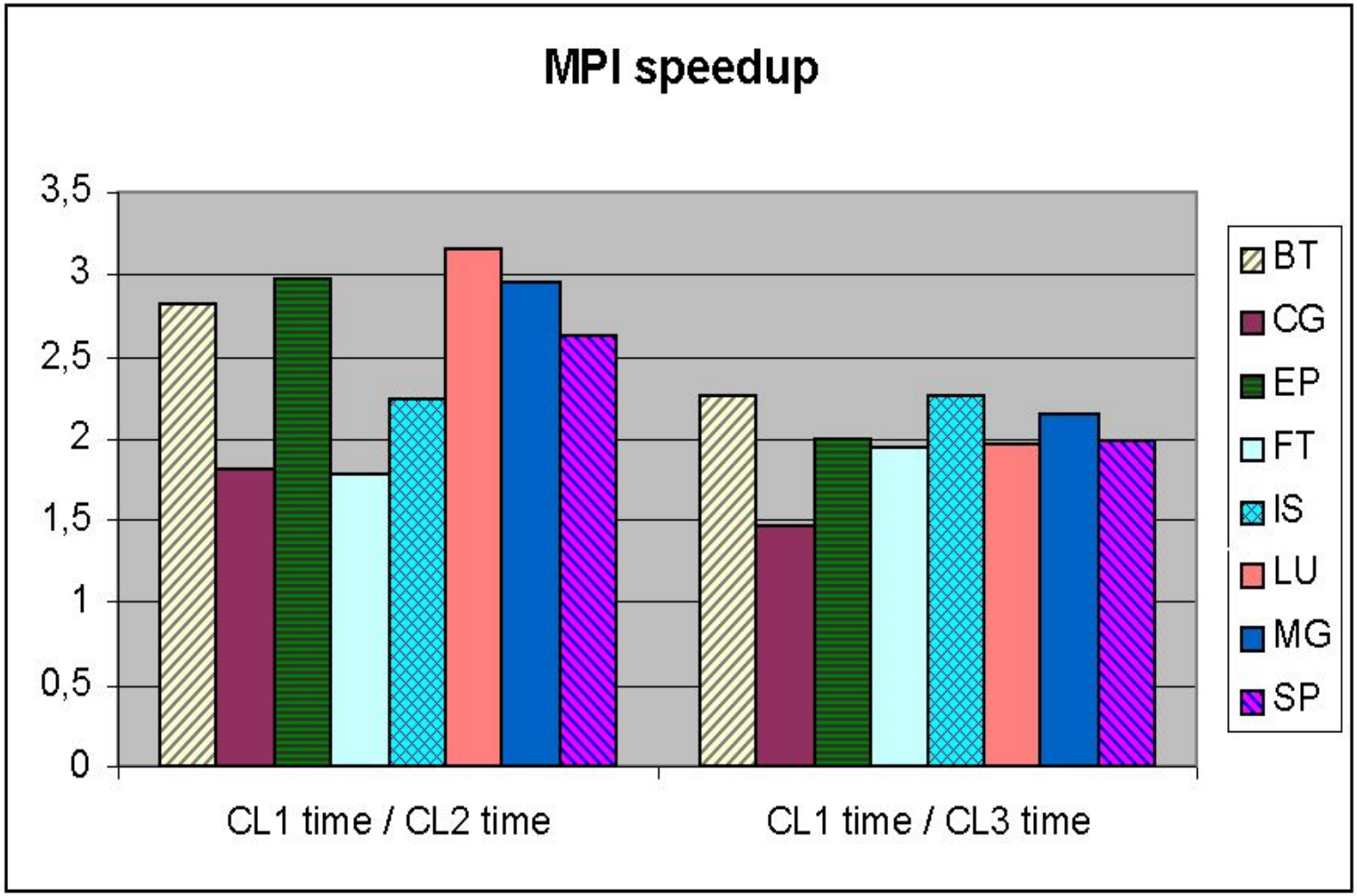
- **BT** 3D Навье-Стокс, метод переменных направлений
- **CG** Оценка наибольшего собственного значения симметричной разреженной матрицы
- **EP** Генерация пар случайных чисел Гаусса
- **FT** Быстрое преобразование Фурье, 3D спектральный метод
- **IS** Параллельная сортировка
- **LU** 3D Навье-Стокс, метод верхней релаксации
- **MG** 3D уравнение Пуассона, метод Multigrid
- **SP** 3D Навье-Стокс, Beam-Warning approximate factorization

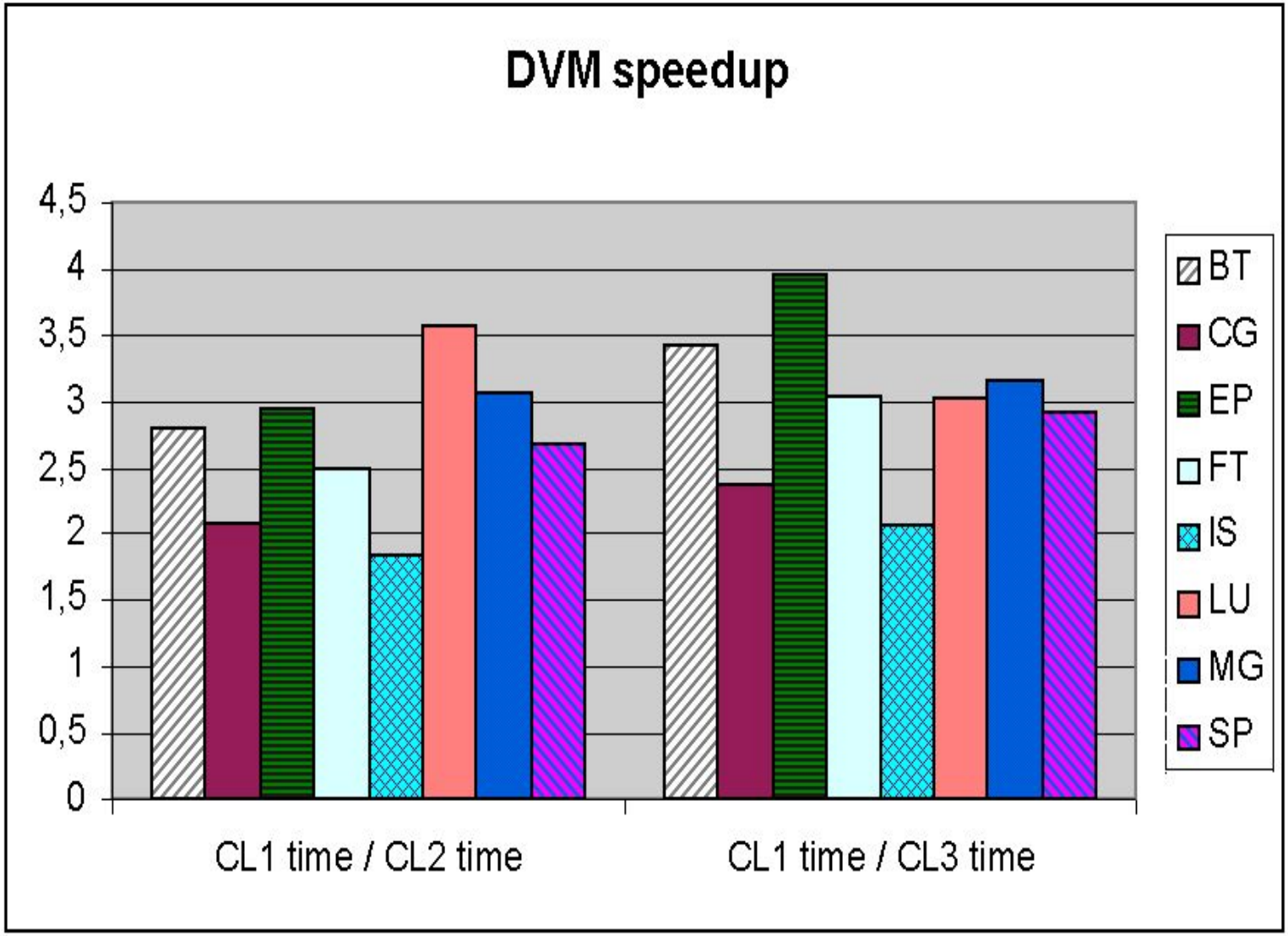
Неоднородный кластер

Неоднородный кластер был промоделирован на машине МВС-1000М с увеличением процессорных времен между последовательными обращениями к MPI функциям.

В следующих диаграммах будет показано изменение времени выполнения MPI и DVM версий тестов NAS (класс C) на следующих конфигурациях:

- CL1 – 128 процессоров со скоростью выполнения P,
- CL2 – 128 процессоров со скоростью выполнения 3P,
- CL3 – 128 процессоров со скоростью выполнения P и 128 процессоров со скоростью выполнения 3P (неоднородный кластер).


MPI speedup
3,5
3
2,5
2
1,5
1
0,5
0
CL1 time / CL2 time
CL1 time / CL3 time
BT
CG
EP
FT
IS
LU
MG
SP


DVM speedup
4,5
4
3,5
3
2,5
2
1,5
1
0,5
0
CL1 time / CL2 time
CL1 time / CL3 time
BT
CG
EP
FT
IS
LU
MG
SP

Неоднородность коммуникационной среды

Способы адаптации к медленным коммуникационным каналам:

- сокращение количества обменов => борьба с высокой латентностью

  - использование языковых средств для группировки операций, требующих обмены (редукции, доступ к удаленным элементам) и дублирования вычислений вместо обмена данных

  - автоматический выбор конфигурации решетки виртуальных процессоров и их отображения на физические процессоры для сокращения количества обменов через медленные коммуникационные каналы

- сокращение объема передаваемой информации посредством использования языковых средств дублирования вычислений вместо обмена данных и автоматической упаковки сообщений => борьба с низкой пропускной способностью

- сокращение вычислений, распределяемых на физические процессоры, связанные между собой медленными коммуникационными каналами => балансировка общей загрузки процессоров