

Управление фрагментацией суперкомпьютерных ресурсов для повышения быстродействия вычислений

Баранов А.В, Брагин К.А.

НИЦ "Курчатовский Институт"
Отделение суперкомпьютерных систем и параллельных
вычислений

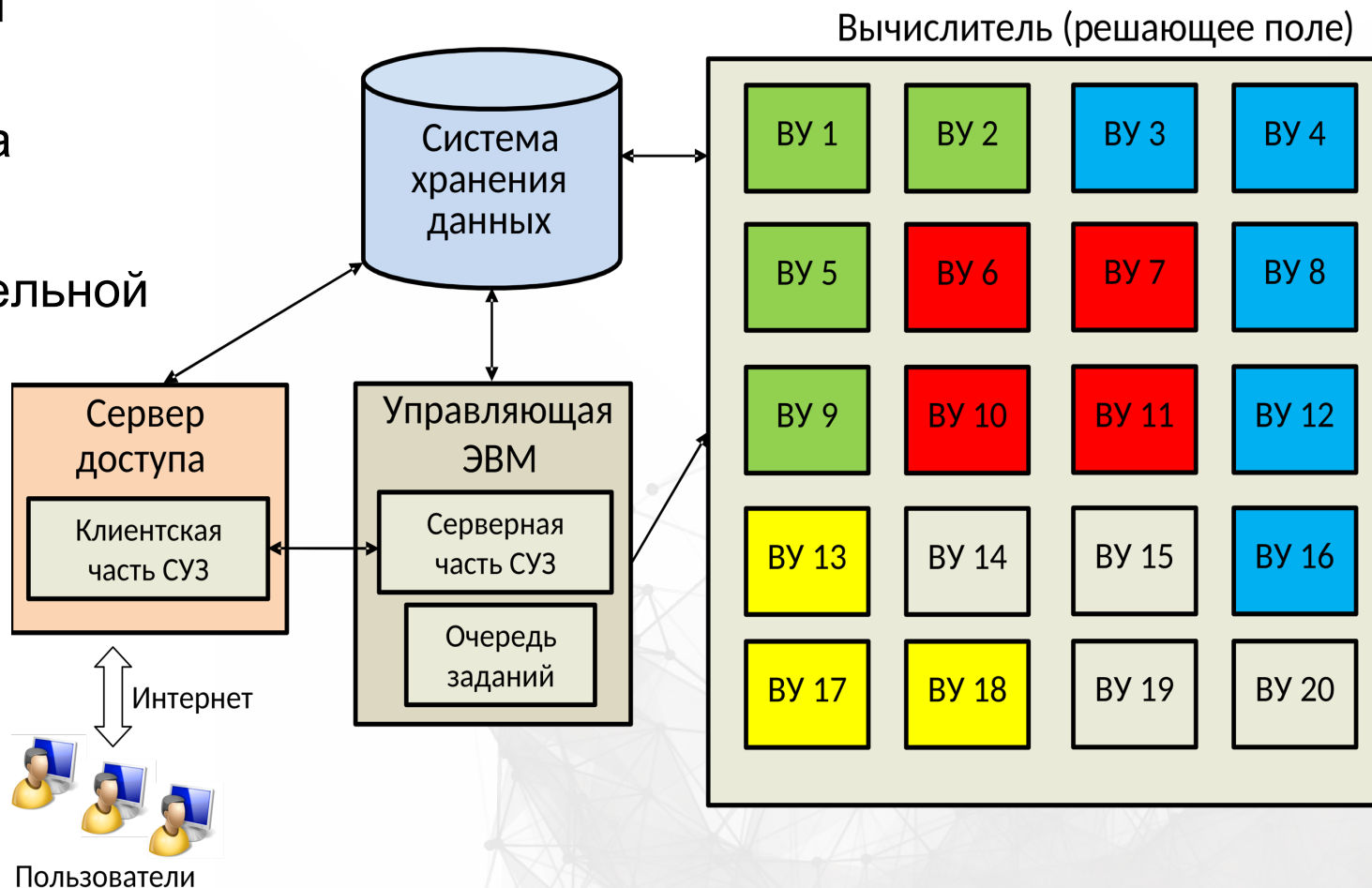
Система коллективного пользования

Задание (Job) – специальный информационный объект:

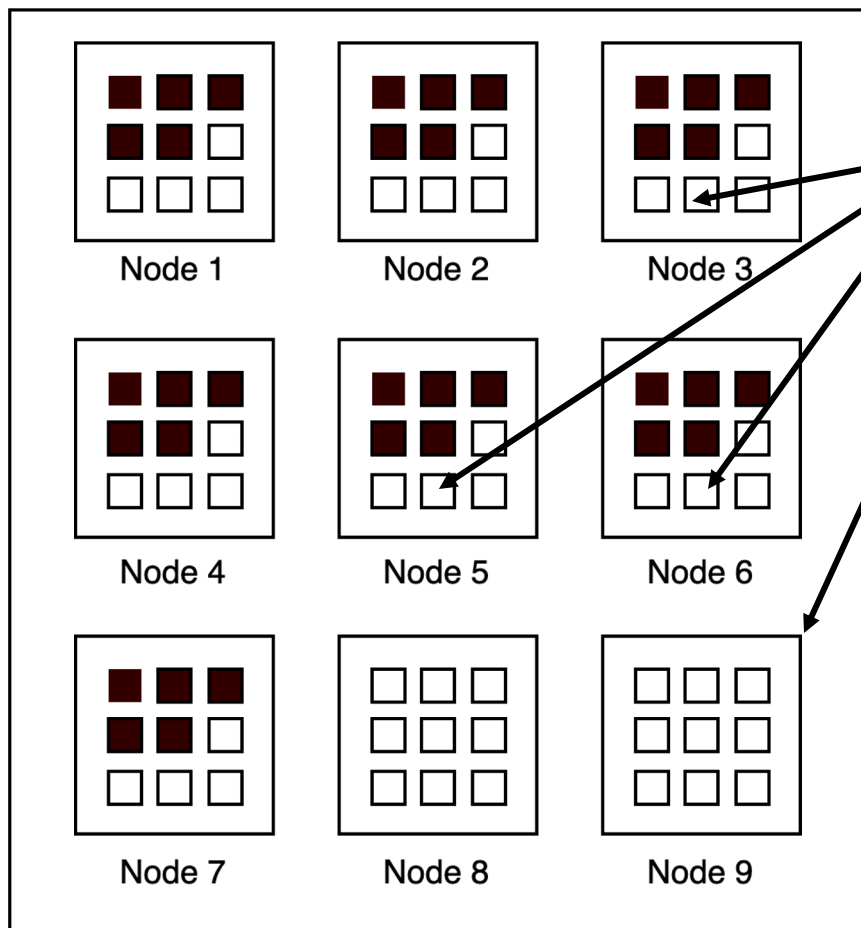
- параллельная программа
- требования к ресурсам
- входные данные параллельной программы



СУПЗ



Фрагментация ресурсов



Фрагментация – наличие неиспользуемых ресурсов вычислительной системы.

- внутренняя фрагментация возникает при простаивающих ядрах отдельного узла
- внешняя фрагментация возникает при наличии свободных узлов

Фрагментация считается **негативным** фактором снижающим загрузку системы.

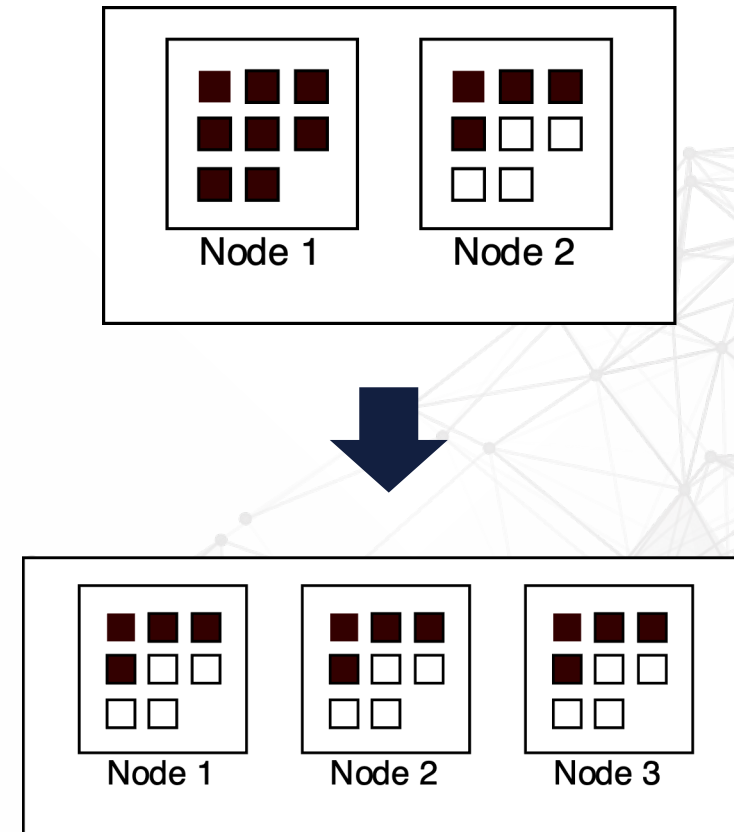
Предположим, что одновременно со снижением загрузки будет снижена и конкуренция за ресурсы внутри узла.

Задание запрашивает M ядер, вычислительная система оборудована K узлами по P ядер, для задания будет выделено R ядер:

$$R = K \cdot P, R \geq M, K \in \mathbb{N}$$

Управление фрагментацией

Искусственное добавление
фрагментации с целью уменьшения
конкуренции процессов задания за
ресурсы и возможным увеличением
быстродействия расчетов



Влияние фрагментации на производительность

Показатели:

Коэффициент
фрагментации

$$K_n = \frac{N + L}{N}$$

Коэффициент
быстродействия

$$K_p = \frac{T_N}{T_{N+L}}$$

Показатель
эффективности

$$E = K_p - K_n, E > 0$$

Естественная
фрагментация F_e , %

$$F_e = 100 - \frac{C_s}{C_N} \times 100$$

Искусственная
фрагментация F_i , %

$$F_i = 100 - \frac{C_s}{C_{N+L}} \times 100$$

N – число узлов, используемых для задачи

L – число искусственно добавленных узлов

T_N – время расчёта на N узлах

T_{N+L} – время расчета на $N+L$ узлах

C_s – свободных ядер на узле

C_N – общее число ядер на узле

C_{N+L} – общее число ядер на $N+L$ узлах

Методика тестирования

NAS Parallel Benchmarks (NPB) — это набор программ, разработанных для оценки производительности суперкомпьютеров. Бенчмарки взяты из задач вычислительной гидродинамики (CFD) и включают пять базовых алгоритмов и три псевдо-приложения.

Базовые алгоритмы

- IS – Целочисленная сортировка
- EP - Генерация независимых нормально распределённых случайных величин;
- CG - Приближение к наименьшему собственному значению большой разреженной симметричной положительно-определённой матрицы;
- MG - Аппроксимация решения трёхмерного дискретного уравнения Пуассона;
- FT - Решение трёхмерного уравнения в частных производных при помощи Быстрого преобразования Фурье;

Псевдо-приложения для решения решения синтетическую систему нелинейных дифференциальных уравнений в частных производных

- BT - методом блочной трёхдиагональная схема с методом переменных направлений
- SP – методом скалярной пятидиагональная схема
- LU – при помощи симметричного метода Гаусса — Зейделя

Вычислительная система для проведения экспериментов



Эксперименты проводились на разделе Cascade Lake суперкомпьютера MBC-10П ОП, содержащим 101 узлов с характеристиками:

- 2 Intel Xeon Gold 6248R процессора;
- 48 физических, 96 виртуальных ядра на каждом узле;
- 192 GB of RAM;
- Intel Omni-Path interconnect.

Оценка влияния фрагментации

Случаи для которых $E > 0$

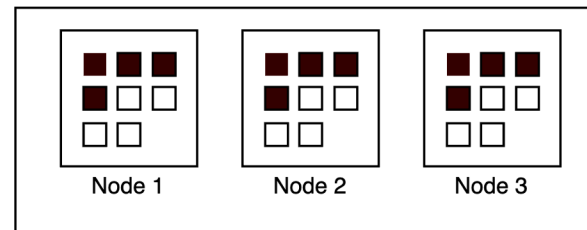
Тест	Узлов (ядер) без ИФ	Естественная фрагментация, %	Время выполнения, с	Узлов (ядер) с ИФ	Искусственная фрагментация, %	Время выполнения , с	Ускорение K_p	Коэф. фгар. K_n
CG	11(528)	3,03	317,62	12(576)	8,33	288,10	1,102	1,091
FT	11(528)	3,03	198,90	12(576)	8,33	174,14	1,142	1,091
FT	11(528)	3,03	198,90	14(672)	21,43	155,41	1,280	1,273
MG	3(144)	11,11	16,52	4(192)	25,00	11,59	1,426	1,333
MG	3(144)	11,11	16,52	5(240)	40,00	9,54	1,732	1,667
MG	6(288)	11,11	8,21	7(336)	14,29	6,90	1,191	1,167
MG	6(288)	11,11	8,21	8(384)	25,00	6,03	1,363	1,333
MG	11(528)	3,03	35,90	12(576)	8,33	31,47	1,141	1,091
MG	11(528)	3,03	35,90	13(624)	15,38	29,07	1,235	1,182
MG	11(528)	3,03	35,90	14(672)	21,43	25,91	1,386	1,273
MG	11(528)	3,03	35,90	15(720)	26,67	24,96	1,438	1,364
MG	11(528)	3,03	35,90	16(768)	31,25	22,94	1,565	1,455
SP	6(288)	11,11	111,39	7(336)	14,29	93,48	1,192	1,167

Управление фрагментацией для потока заданий

1. Выделение дополнительных узлов из множества свободных в настоящий момент **увеличивает время** ожидания других заданий в очереди
2. Выделение свободных узлов **ограничивает** число выполняемых одновременно пользовательских программы на всей вычислительной системе
3. Искусственное введение внутренней фрагментации может привести к появлению внешней

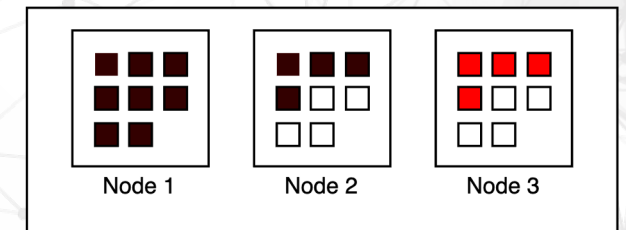
Очередь с применением метода:

- Задача 1
- **Задача 2**
- Задача 3
- **Задача 4**



Очередь без применения метода:

- Задача 1
- **Задача 2**
- Задача 3
- **Задача 4**



Имитационное моделирование

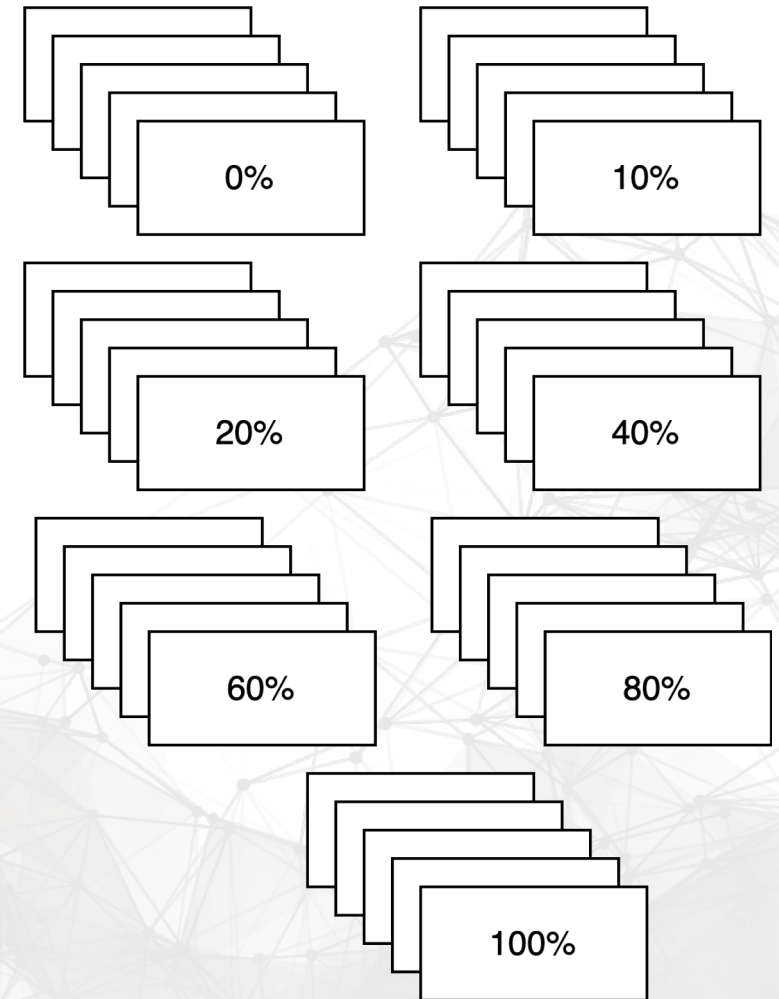
Входные данные:

- Смесь из 100 пользовательских заданий
- 7 групп смесей
- Процентное содержание задач с применением метода от 0% до 100%
- По 5 запусков в каждой группе
- 3500 записей пользовательских заданий

Система Alea

- Предоставляет различные алгоритмы планирования
- Позволяет проводить моделирование систем с различным числом вычислительных узлов
- Детерминированная система

Klusáček, D., Soysal, M., Suter, F. (2020). Alea – Complex Job Scheduling Simulator. In: Wyrzykowski, R., Deelman, E., Dongarra, J., Karczewski, K. (eds) Parallel Processing and Applied Mathematics. PPAM 2019. Lecture Notes in Computer Science(), vol 12044. Springer, Cham. https://doi.org/10.1007/978-3-030-43222-5_19.



Имитационное моделирование

Показатели:

Среднее время
обработки смеси

$$T_c(P) = \frac{1}{k} \sum_{j=1}^k T_{c,j}(P)$$

Ускорение для заданной
смеси с процентом
фрагментированных
заданий P

$$A(P) = \frac{T_c(P)}{T_c(0)}$$

Число узлов в системе

N

Внешняя фрагментация
системы

$$F = \frac{N_f}{N}$$

P – процент фрагментированных заданий

$T_{c,j}(P)$ – время обработки смеси s для j -го запуска

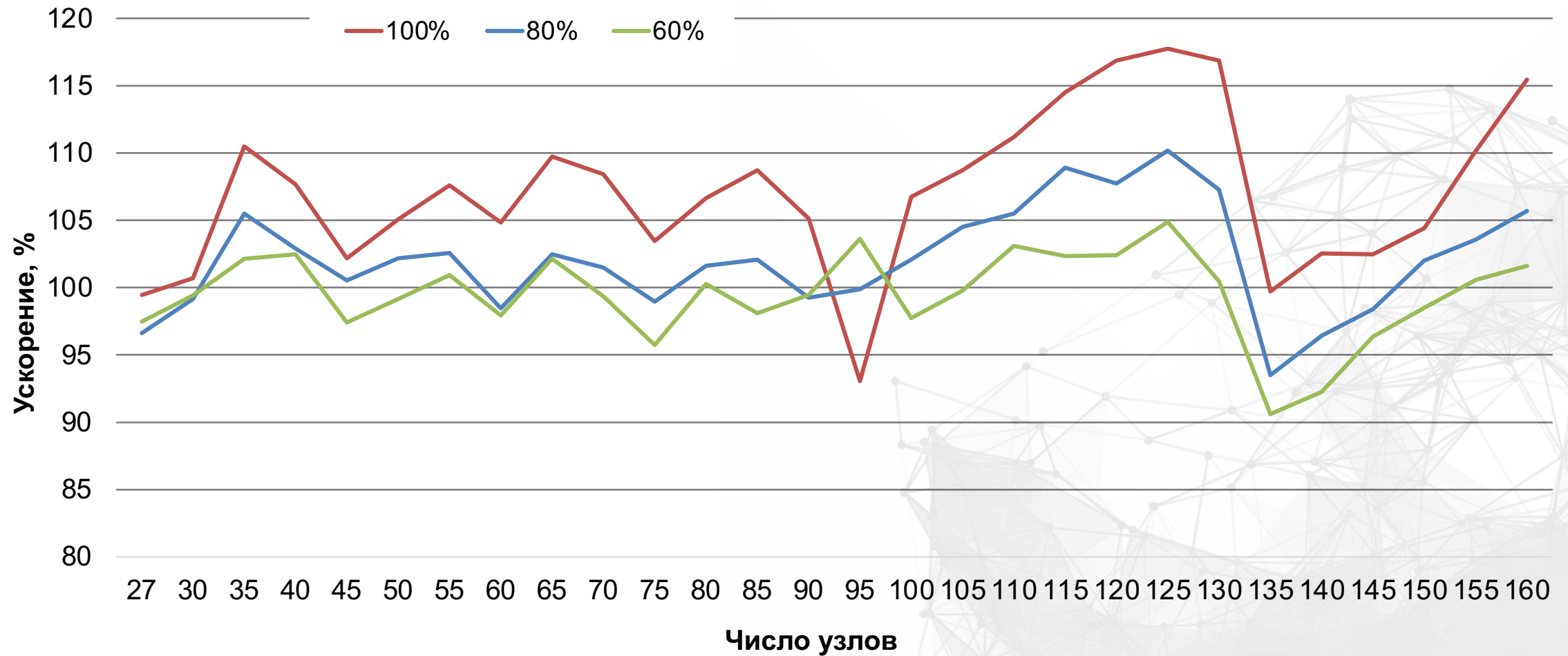
k – число повторений для каждой смеси

$T_c(0)$ – среднее время обработки смесей без фрагментированных заданий

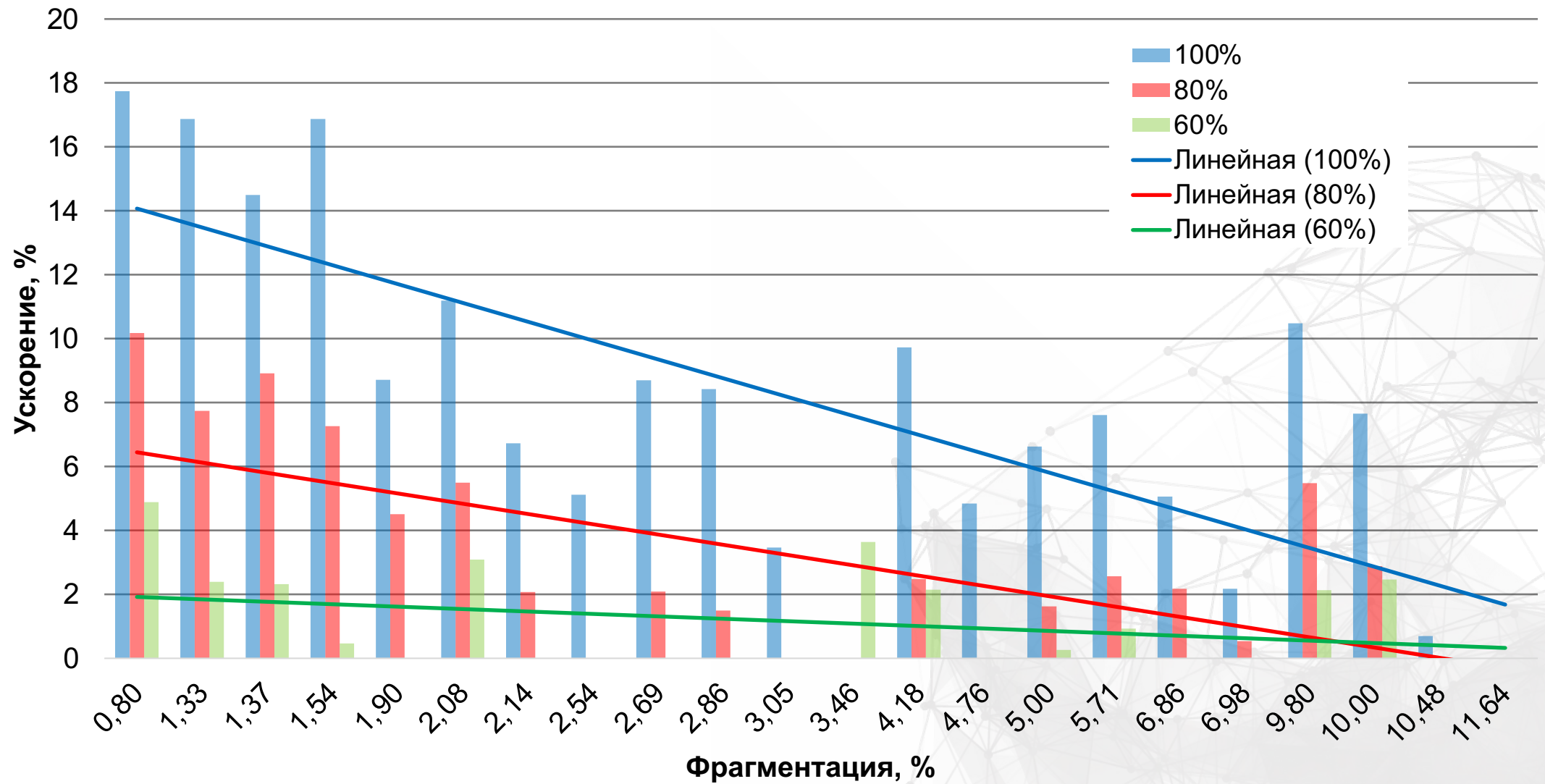
N – общее число узлов системы,

N_f – число узлов свободных в настоящий момент.

Зависимость ускорения от числа узлов системы



Зависимость ускорения от фрагментации системы



Выводы и перспективы

Выводы:

- 1) Искусственная фрагментация позволяет получить статистически значимое повышение скорости обработки входного потока заданий
- 2) Предложен критерий оценки эффективности и определены граничные значения области применимости
- 3) Для значительного числа пользовательских программ на примере набора NPВ предложенный критерий выполняется

Перспективы:

- 1) Оценить влияние точечного распределения процессов программы на различные ядра в рамках одного узла
- 2) Оценить влияние точечного распределения процессов программы на различные ядра в рамках нескольких узлов

Благодарю за внимание